Analyse canonique

Classification

Composantes principales

Corrélation

Espérance conditionnelle

Probabilités analyse des données et Statistique

Estimation

Monte-Carlo

Régression

Tests

Vraisemblance

Gilbert SAPORTA



Gilbert SAPORTA
Professeur au Conservatoire National
des Arts et Métiers

PROBABILITÉS
ANALYSE DES DONNÉES
ET STATISTIQUE

2[e] édition révisée et augmentée

2006

Editions TECHNIP 27 rue Ginoux, 75737 PARIS Cedex 15, FRANCE

CHEZ LE MÊME ÉDITEUR

- Approche pragmatique de la classification
  J.P. NAKACHE, J. CONFAIS
- Data mining et statistique décisionnelle
  S. TUFFÉRY
- Statistique explicative appliquée
  J.P. NAKACHE, J. CONFAIS
- Les techniques de sondage, nouvelle édition
  P. ARDILLY
- Analyse statistique des données spatiales
  J.-J. DROESBEKE, M. LEJEUNE, G. SAPORTA, Eds.
- Modèles statistiques pour données qualitatives
  J.-J. DROESBEKE, M. LEJEUNE, G. SAPORTA, Eds.
- Plans d'expériences. Applications à l'entreprise
  J.-J. DROESBEKE, J. FINE, G. SAPORTA, Eds.
- Méthodes bayésiennes en statistique
  J.-J. DROESBEKE, J. FINE, G. SAPORTA, Eds.
- La régression PLS. Théorie et pratique
  M. TENENHAUS
- Probabilités. Exercices corrigés
  D. GHORBANZADEH

Imprimé en France

ISBN 2-7108-0814-5

Avant-propos

La précédente édition de cet ouvrage a été publiée en 1990. Nous évoquions alors les évolutions de la statistique de la décennie passée. Depuis lors, les progrès de l'informatique n'ont cessé, permettant d'une part l'utilisation de nouvelles méthodes fondées sur des calculs intensifs (simulation, méthodes non-paramétriques et algorithmiques), et d'autre part le traitement de données en masse qui a donné lieu à l'émergence du « data mining » ou « fouille de données ». Les logiciels de calcul statistique n'ont cessé de se perfectionner et de se diffuser à tel point que des méthodes complexes sont employées de façon routinière sans pour cela que l'utilisateur les domine toujours.

Cette nouvelle édition prend en compte ces évolutions. Outre une mise à jour de certains exemples, les principaux développements concernent les méthodes de Monte Carlo, l'estimation non paramétrique, la modélisation prédictive avec l'introduction des méthodes de régression en présence de multicolinéarité, la régression logistique, les SVM et les techniques d'apprentissage. Nous avons également rajouté deux chapitres consacrés aux deux grandes méthodologies de recueil des données : sondages et plans d'expériences. Ce livre a pour but de donner aux étudiants et aux praticiens les outils nécessaires pour appliquer correctement les méthodes statistiques. La plupart des résultats sont démontrés, sauf certains pour lesquels les preuves trop techniques auraient alourdi ce livre. Les 21 chapitres sont regroupés en cinq parties :

La première « outils probabilistes » donne les bases nécessaires à l'inférence classique. L'approche probabiliste permet de prendre en compte le fait que notre univers n'est pas déterministe et que les données dont on dispose ne sont pas parfaites. La deuxième partie intitulée « statistique exploratoire » regroupe les outils de description non-probabilistes des données, allant de la statistique descriptive unidimensionnelle à ce que l'on appelle « analyse des données » en un sens restreint qui selon nous ne se limite pas aux méthodes dérivées de l'analyse en composantes principales et de la classification : pour nous le but de la statistique est d'analyser des données . . . La troisième partie « statistique inférentielle » est consacrée classiquement à l'estimation et aux tests. La quatrième partie « modèles prédictifs » regroupe les techniques de régression au sens large où on cherche un modèle reliant une réponse Y à des prédicteurs X_j. La cinquième partie concerne « le recueil des données » par sondages ou expérimentation. Le recueil des données constitue un préalable à l'analyse ; le placer en dernière partie peut sembler

illogique, mais le fait est que la collecte des données ne peut se concevoir sans en connaître l'usage ultérieur, ce qui nécessite la compréhension de l'estimation et de la modélisation.

Je remercie enfin tous ceux qui ont contribué à un titre ou à un autre à la réalisation de cet ouvrage, ainsi que les Éditions Technip pour leur patience et le soin apporté à sa réalisation.

Je remercie enfin tous ceux qui ont contribué

Gilbert Saporta
(mars 2006)

Table des matières

Deuxième partie : Statistique exploratoire

Troisième partie : Statistique inférentielle

Cinquième partie : Recueil des données

Introduction

Les méthodes statistiques sont aujourd'hui utilisées dans presque tous les secteurs de l'activité humaine et font partie des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du biologiste, de l'informaticien ... Parmi les innombrables applications citons dans le domaine industriel : la fiabilité des matériels, le contrôle de qualité, l'analyse des résultats de mesure et leur planification, la prévision, et dans le domaine de l'économie et des sciences de l'homme : les modèles économétriques, les sondages, les enquêtes d'opinion, les études quantitatives de marché, etc.

Nous allons tenter de préciser dans les paragraphes suivants les notions fondamentales de la statistique et les rapports qu'elle entretient avec la théorie des probabilités ainsi que ce qu'on entend par démarche statistique.

LA STATISTIQUE, LES STATISTIQUES ET LE CALCUL DES PROBABILITÉS

Selon la définition de l'*Encyclopedia Universalis :* « Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation ».

Ainsi le relevé des débits journaliers d'une rivière de 1971 à 1983 constitue une statistique tandis que faire de la statistique sur ces données consisterait par exemple, à tracer des graphiques mettant en évidence la périodicité du phénomène, à calculer un débit moyen ou à prévoir la valeur maximale de la crue annuelle.

Individus et variables

Définitions générales

Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées « variables ».

Ainsi en contrôle de fabrication on prélèvera un ensemble de pièces dans une production homogène et on mesurera leur poids, leur diamètre. En marketing on étudiera les clients

d'une entreprise en les décrivant par leurs caractéristiques socio-démographiques et leurs achats passés.

La notion fondamentale en statistique est celle de groupe ou d'ensemble d'objets équivalents que l'on appelle ***population***. Ce terme hérité des premières applications de la statistique à la démographie est employé pour désigner toute collection d'objets à étudier ayant des propriétés communes. Ces objets sont appelés des ***individus*** ou ***unités statistiques***.

La statistique traite des propriétés des populations ou de sous-populations plus que de celles d'individus particuliers :

Généralement la population à étudier est trop vaste pour pouvoir être observée exhaustivement : c'est évidemment le cas lorsque la population est infinie : par exemple l'ensemble de toutes les pièces métalliques que pourrait sortir une machine dans des conditions de fabrication déterminées, mais c'est aussi le cas lorsque les observations sont coûteuses (contrôle destructif entre autres).

L'étude de tous les individus d'une population finie s'appelle un ***recensement***. Lorsque l'on n'observe qu'une partie de la population on parle de ***sondage***, la partie étudiée s'appellant ***l'échantillon***.

Chaque individu d'une population est décrit par un ensemble de caractéristiques appelées ***variables*** ou caractères. Ces variables peuvent être classées selon leur nature :

– variables ***quantitatives*** ou numériques : par exemple taille, poids, volume, s'expriment par des nombres réels sur lesquels les opérations arithmétiques courantes (somme, moyenne, . . .) ont un sens. Certaines peuvent être ***discrètes*** (nombre fini ou dénombrable de valeurs) : nombre de défauts d'une pièce, de véhicules passant en une heure à un péage, etc. ou ***continues*** si toutes les valeurs d'un intervalle de $\mathbb{R}$ sont acceptables.
– variables ***qualitatives*** s'exprimant par l'appartenance à une ***catégorie*** ou ***modalité*** d'un ensemble fini. Certaines sont purement ***nominales*** : par exemple type de traitement thermique subi par un alliage, catégorie socio-professionnelle d'un actif (ouvrier, cadre, employé . . .), d'autres sont ***ordinales*** lorsque l'ensemble des catégories est muni d'un ordre total ; par exemple : très résistant, assez résistant, peu résistant.

Le concept clé en statistique est la ***variabilité*** qui signifie que des individus en apparence semblables peuvent prendre des valeurs différentes : ainsi un processus industriel de fabrication ne fournit jamais des caractéristiques parfaitement constantes.

L'analyse statistique est pour l'essentiel une étude de la variabilité : on peut en tenir compte pour prévoir de façon probabiliste le comportement d'individus non encore observés, chercher à la réduire ou « l'expliquer » à l'aide de variables extérieures, ou chercher à l'augmenter dans le but de distinguer le mieux possible les individus entre eux.

Tableaux de données

On présente usuellement sous forme de tableau à n lignes, les données recueillies sur n individus. Lorsque l'on observe uniquement des variables numériques le tableau a la forme d'une matrice à n lignes et p colonnes de terme général x_i^j :

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 2 \\ . \\ i \\ . \\ . \\ . \\ n \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{x}^1 & \mathbf{x}^2 & \mathbf{x}^j & \mathbf{x}^p \end{array} \\ \left[\begin{array}{ccc} & . & \\ & . & \\ & . & \\ \dots & x_i^j & \\ & . & \\ & . & \\ & . & \\ & . & \end{array} \right] \end{array}$$

Lorsque les variables sont toutes qualitatives, le tableau où x_i^j désigne le numéro de la catégorie de la variable $\mathcal{X}^j$ à laquelle appartient l'individu i est le tableau des codages réduits. Les numéros des modalités étant arbitraires, on lui associera le tableau disjonctif à $m_1 + m_2 + \ldots + m_p$ colonnes constitué de la façon suivante :

A toute variable à m_j catégories on substitue un ensemble de m_j variables valant 0 ou 1 (les indicatrices des catégories). Ainsi au tableau 5×3 des observations sur 5 individus de 3 variables à 2, 3 et 2 modalités respectivement :

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 1 \\ 2 & 1 & 2 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

correspond le ***tableau disjonctif*** à 5 lignes et 7 colonnes suivant :

$$\left[\begin{array}{cc:ccc:cc} 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right]$$

Mentionnons enfin les ***tableaux de contingence*** ou tableaux croisés qui résultent d'un premier traitement et fournissent la ventilation de n individus selon deux variables qualitatives à m_1 et m_2 modalités :

$$\mathbf{N} = \begin{array}{c} \\ 1 \\ 2 \\ . \\ i \\ . \\ m_1 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & j & . & m_2 \end{array} \\ \left[\begin{array}{ccc} & . & \\ & . & \\ & . & \\ \dots & n_{ij} & \dots \\ & . & \\ & . & \end{array} \right] \end{array}$$

où n_{ij} est le nombre d'individus appartenant simultanément aux catégories i et j des deux variables.

Statistique et probabilités

La théorie des probabilités est une branche des mathématiques qui traite des propriétés de certaines structures modélisant des phénomènes où le « hasard » intervient. En tant que théorie mathématique abstraite, elle repose sur une axiomatique et se développe de façon autonome par rapport à la réalité physique. Seuls les noms des concepts utilisés (événements, variables . . .) renvoient à l'expérience.

La théorie des probabilités permet de modéliser efficacement certains phénomènes aléatoires et d'en faire l'étude théorique.

Quels sont ses liens avec la statistique qui repose plutôt sur l'observation de phénomènes concrets ? On peut en voir schématiquement trois : tout d'abord les données observées sont souvent imprécises, entachées d'erreur. Le modèle probabiliste permet alors de représenter comme des variables aléatoires les déviations entre « vraies » valeurs et valeurs observées.

Deuxièmement on constate souvent que la répartition statistique d'une variable au sein d'une population est voisine de modèles mathématiques proposés par le calcul des probabilités (lois de probabilité).

Enfin et c'est à notre avis le rôle le plus important du calcul des probabilités, les échantillons d'individus observés sont la plupart du temps tirés au hasard dans la population, ceci pour assurer mathématiquement leur représentativité : si le tirage est fait de manière équiprobable chaque individu de la population a une probabilité constante et bien définie d'appartenir à l'échantillon. Les caractéristiques observées sur l'échantillon deviennent, grâce à ce tirage au sort, des variables aléatoires et le calcul des probabilités permet d'étudier leurs répartitions. Mentionnons ici les méthodes de validation par rééchantillonnage (bootstrap, validation croisée) qui consistent à re-tirer des observations à l'intérieur de l'échantillon initial.

Il faut bien distinguer ce dernier rôle du calcul des probabilités des deux premiers : dans les premiers cas le calcul des probabilités propose des modèles simplificateurs, éventuellement contestables, du comportement d'un phénomène (par exemple supposer que la durée de vie X d'un composant électronique suit une loi exponentielle $P(X > x) = \exp(-cx)$) ; dans le dernier cas, le calcul des probabilités fournit des théorèmes si le processus d'échantillonnage est respecté : ainsi le théorème central limite permet d'établir que la moyenne $\bar{x}$ d'une variable numérique mesurée sur n individus s'écarte de la moyenne m de la population selon une loi approximativement gaussienne.

Le calcul des probabilités est donc un des outils essentiels de la statistique pour pouvoir extrapoler à la population les résultats constatés sur l'échantillon mais on ne peut y réduire la statistique : à côté du calcul des probabilités, la statistique utilise des mathématiques assez classiques (algèbre linéaire, géométrie euclidienne) et de plus en plus l'informatique, car les calculs à mettre en œuvre nécessitent l'emploi d'ordinateurs : l'informatique a révolutionné la pratique de la statistique en permettant la prise en compte de données multidimensionnelles ainsi que l'exploration rapide par simulation de nombreuses hypothèses.

Ce livre met plus l'accent sur les techniques et la démarche statistiques que sur la théorie des probabilités, conçue ici comme un outil pour la statistique et non comme un objet d'étude en elle-même.

LA DÉMARCHE STATISTIQUE CLASSIQUE

Elle comporte usuellement trois phases : le recueil, l'exploration, l'inférence et la modélisation.

Le recueil des données

En dehors des cas où les données sont déjà disponibles, il est nécessaire de les collecter. Les deux grandes méthodologies sont les sondages et les plans d'expériences.

Les sondages

Essentiellement utilisés dans les sciences humaines, mais également pour obtenir des échantillons dans des bases de données, les techniques de sondages servent à choisir dans une population les unités à interroger ou observer. Le choix des unités se fait en général aléatoirement, mais pas nécessairement avec des probabilités égales pour toutes les unités. L'important est qu'il n'y ait pas d'individus de la population qui aient une probabilité nulle de figurer dans l'échantillon, sinon les résultats risquent d'être ***biaisés*** car l'échantillon ne sera plus ***représentatif***. Les méthodes non-aléatoires sont également souvent utilisées dans les études de marché et d'opinion qui constituent un secteur d'activité important.

Les plans d'expériences

Introduits au début du XXe siècle en agronomie, puis utilisés en recherche industrielle, ils ont pour but de provoquer l'apparition de données selon des conditions expérimentales précises. La théorie des plans d'expériences permet de minimiser le coût de recueil des données en cherchant les expériences les plus efficaces.

Bien qu'employées dans des contextes très différents, ces deux méthodologies ont des points communs : elles cherchent à optimiser le recueil des données. Mais il n'y a pas d'optimum en soi, tout dépend du but recherché. En sondages on cherche à estimer les paramètres d'une population avec une variance minimale en utilisant toutes les informations dont on dispose. Dans les plans d'expériences, on dispose d'un modèle prédictif reliant approximativement une réponse à des facteurs de variabilité : on cherche à déterminer les expériences permettant d'estimer au mieux les paramètres du modèle, ou les prévisions qui en découlent : un plan optimal pour un modèle ne le sera pas pour un autre.

La statistique exploratoire

Son but est de synthétiser, résumer, structurer l'information contenue dans les données. Elle utilise pour cela des représentations des données sous forme de tableaux, de graphiques, d'indicateurs numériques.

Le rôle de la statistique exploratoire est de mettre en évidence des propriétés de l'échantillon et de suggérer des hypothèses. Les modèles probabilistes ne jouent ici qu'un rôle très restreint voire même nul.

Les principales méthodes de l'analyse exploratoire se séparent en deux groupes : Aprés une phase de description variable par variable, puis par couples de variables (la ***statistique descriptive*** classique) l'***analyse des données*** au sens français restreint, exploite le caractére multidimensionnel des observations au mogen de :

- méthodes de ***classification*** visant à réduire la taille de l'ensemble des individus en formant des groupes homogènes;
- méthodes factorielles qui cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques. Selon que l'on travaille avec un tableau de variables numériques ou qualitatives on utilisera l'***analyse en composantes principales*** ou l'***analyse des correspondances***. Les liens entre groupes de variables peuvent être traités par l'***analyse canonique***.

La statistique inférentielle

Son but est d'étendre les propriétés constatées sur l'échantillon à la population toute entière et de valider ou d'infirmer des hypothèses *a priori* ou formulées après une phase exploratoire. Le calcul des probabilités joue souvent un rôle fondamental.

Donnons ici quelques exemples élémentaires.

Estimation d'une moyenne

Une même grandeur est mesurée n fois de suite par un même observateur, l'imprécision de l'instrument de mesure et d'autres facteurs rendent fluctuantes ces mesures et on obtient n valeurs différentes $x_1, x_2, \ldots, x_n$. Comment déterminer la vraie valeur m ? On peut admettre que ces valeurs constituent des observations ou ***réalisations*** indépendantes d'une variable X de moyenne théorique m (***espérance mathématique***) si il n'y a pas d'erreurs systématiques. La loi des grands nombres montre alors que la moyenne $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$ de l'échantillon constitue une bonne approximation de m; $\bar{x}$ est une ***estimation*** de m.

L'échantillon ayant été « tiré au hasard » la valeur constatée $\bar{x}$ n'est qu'une de celles que l'on aurait pu trouver : c'est donc une variable aléatoire qui aurait pu fournir une autre valeur si on avait répété l'expérience dans les mêmes conditions.

Si n est assez grand le calcul des probabilités fournit avec une grande précision la loi de répartition des valeurs possibles de $\bar{x}$ autour de m et on pourrait en déduire si m était connu un intervalle du type $[m - \Delta m, m + \Delta m]$ ayant une probabilité fixée, disons 95 %, de contenir $\bar{x}$. Connaissant une observation $\bar{x}$ on inverse alors la problématique et on peut en déduire une ***fourchette*** ou ***intervalle de confiance*** pour la vraie valeur m.

Vérification d'une hypothèse ou test

Le cas suivant est classique en contrôle de qualité. Un client commande à son fournisseur des lots de pièces dont la qualité est spécifiée par contrat : le fournisseur s'engage à respecter un taux de pièces défectueuses inférieur à 4 %. Avant de livrer, le fournisseur effectue un

contrôle sur 50 pièces et en trouve trois défectueuses soit 6 % : doit-il livrer quand même au risque de se faire refuser la marchandise ?

Le raisonnement est alors le suivant : si le taux théorique de défectueux est de 4 % quelles sont les chances d'observer un tel nombre de défectueux ? Le calcul des probabilités montre alors qu'il y a une probabilité voisine de 0.32 d'observer trois pièces défectueuses ou plus (loi binomiale $\mathcal{B}(50 ; 0.04)$). Cette probabilité étant assez forte, l'événement constaté paraît donc normal au fournisseur et ne semble pas de nature à remettre en cause l'hypothèse formulée. Mais le client serait-il d'accord ? . . . Il faut alors calculer le risque d'un refus par le client.

Dans ces deux cas le raisonnement procède du même schéma :
- l'échantillon est tiré au hasard dans une population plus vaste ;
- le calcul des probabilités permet ensuite de préciser les caractéristiques de l'ensemble des échantillons que l'on aurait pu obtenir par le même procédé, c'est l'étude des ***distributions d'échantillonnage*** ;
- on inverse les conclusions de la phase précédente pour en déduire la structure vraisemblable de la population dont est issu l'échantillon observé. C'est la phase inférentielle.

On ne manquera pas de constater la similitude de cette démarche statistique avec la démarche scientifique habituelle : observation, hypothèses, vérification.

L'avènement des ordinateurs et le développement du calcul statistique permettent dans une certaine mesure de s'affranchir de modèles probabilistes souvent illusoires car choisis pour leur relative simplicité mathématique mais pas toujours adaptés aux données. Les méthodes de rééchantillonnage renouvellent la problématique de l'inférence en n'utilisant que les données observées.

La modélisation et la prévision statistique

La modélisation consiste généralement à rechercher une relation approximative entre une variable et plusieurs autres, la forme de cette relation étant le plus souvent linéaire. Lorsque la variable à « expliquer » ou à prévoir est numérique ainsi que les variables explicatives, on parle de ***régression linéaire***, si certaines variables explicatives sont qualitatives le ***modèle linéaire général*** en est une extension.

Lorsque l'on cherche à prévoir une variable qualitative (appartenance à une catégorie) on utilisera une méthode de ***discrimination***.

STATISTIQUE ET « DATA MINING »

L'émergence d'immenses bases de données, souvent recueillies automatiquement, en particulier dans le fonctionnement des entreprises, a fait apparaître de nouvelles problématiques, différentes de celles exposées précédemment. Il ne s'agit plus tant de découvrir ou d'estimer des modèles de la réalité (démarche scientifique) mais de donner des réponses à des questions opérationnelles comme : à quelles adresses d'un fichier dois-je envoyer une

publicité pour obtenir un taux de retour maximal, à qui dois-je accorder un crédit pour minimiser le risque de perte ?

La statistique n'est plus alors un auxiliaire de la science mais aussi un outil pour l'action.

Le « data mining » que l'on peut traduire par « ***fouille de données*** » est apparu au milieu des années 1990 comme une nouvelle discipline à l'interface de la statistique et des technologies de l'information : bases de données, intelligence artificielle, apprentissage automatique (*machine learning*).

David Hand (1998) en donne la définition suivante : *« Data Mining consists in the discovery of interesting, unexpected, or valuable structures in large data sets »*. La métaphore qui consiste à considérer les grandes bases de données comme des gisements d'où l'on peut extraire des pépites à l'aide d'outils spécifiques n'est pas nouvelle. Dès les années 1970 Jean-Paul Benzécri n'assignait-il pas le même objectif à l'analyse des données ? : *« L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature »*.

Tel M. Jourdain, les statisticiens faisaient donc du data mining sans le savoir.

« Data Mining » versus modélisation statistique

La notion de modèle en fouille de données prend un sens particulier : un modèle est une relation entre des variables exprimable sous une forme analytique ou algorithmique qui ne provient pas d'une théorie mais est issu de l'exploration des données et réalise un bon ajustement. Ainsi il est courant d'explorer différents modèles (linéaires, non-linéaires) en faisant varier les paramètres (nombre de couches dans un réseau de neurones, noyau pour des SVM etc.) jusqu'à obtenir les meilleures prédictions. On est très loin de la démarche usuelle de modélisation, mais plutôt dans une optique pragmatique où il ne s'agit pas forcément de comprendre mais de prévoir du mieux possible.

La démarche n'est pas pour autant du pur empirisme et se fonde sur la théorie de l'apprentissage statistique : un modèle réalise un compromis entre sa capacité à rendre compte des données d'apprentissage et sa capacité de généralisation à de nouvelles données.

L'inférence statistique classique a été développée pour traiter des « petits » échantillons. En présence de très grandes bases de données le paradoxe est que tout devient significatif : par exemple, pour un million d'individus, l'hypothèse d'indépendance entre deux variables sera rejetée au risque 5 % si le coefficient de corrélation linéaire est supérieur en valeur absolue à 0.002, ce qui est sans intérêt. L'inférence classique ne fonctionne plus et la fouille des grandes bases de données amène à repenser la notion de test : le choix d'un modèle se fait en fonction de ses performances sur d'autres données que celles qui ont servi à le choisir et le caler, d'où l'emploi de méthodes de validation croisée ou de mesures de capacité de type dimension de Vapnik-Cervonenkis. En outre en Data Mining, on analyse des données recueillies à d'autres fins : c'est une ***analyse secondaire*** destinée à valoriser des bases de données déjà constituées : on ne se préoccupe plus de collecter des données de manière efficace. L'échantillonnage ne perd cependant pas ses droits dans la phase de validation car il est souvent préférable de travailler sur une partie de la base que sur la totalité.

Plutôt que d'opposer data mining et statistique, il vaut mieux considérer que le data mining représente la branche de la statistique consacrée à l'exploitation des grandes bases de

données. Si de nouvelles méthodes ont vu le jour en dehors du monde des statisticiens, il n'en reste pas moins que ces méthodes relèvent de la statistique au sens large « recueil, traitement, interprétation de données » et que l' « esprit statistique » imprégné des notions de marge d'erreur, de risque, d'incertain, reste indispensable pour en relativiser les conclusions.

Le praticien de la statistique doit donc être non seulement à l'interface avec les disciplines d'application, mais aussi dominer les outils informatiques de son temps.

1 Le modèle probabiliste

En tant que théorie mathématique, la théorie des probabilités n'a pas à être justifiée : une fois ses axiomes posés, elle se développe de façon autonome par rapport à la réalité concrète.

Il en va autrement lorsque l'on cherche à appliquer le calcul des probabilités : on ne peut alors éluder la question de la nature de la probabilité et de la validité du modèle probabiliste. Après trois paragraphes consacrés à un exposé simple[1] de la théorie on se proposera de donner quelques éléments de réflexion sur le concept de probabilité.

1.1 ESPACE PROBABILISABLE

On expose ici la formalisation d'une expérience où intervient le « hasard ».

1.1.1 Expérience aléatoire et événements

Une ***expérience*** est qualifiée d'***aléatoire*** si l'on ne peut prévoir par avance son résultat et si, répétée dans des conditions identiques, elle peut (on aurait pu s'il s'agit d'une expérience par nature unique) donner lieu à des résultats différents.

On représente le résultat de cette expérience comme un élément ω de l'ensemble Ω de tous les résultats possibles : Ω est appelé ***l'ensemble fondamental*** ou encore l'univers des possibles.

Ainsi à l'expérience aléatoire qui consiste à lancer deux dés, on peut associer l'ensemble $\Omega = \{(1.1), (1.2), (1.3)\ldots\}$ à 36 éléments.

Il convient de noter ici que l'ensemble Ω ne se déduit pas de manière unique de l'expérience mais dépend de l'usage qui doit être fait des résultats : ainsi, si l'on convient une fois pour toutes qu'on ne retiendra de l'expérience des deux dés que la somme des points affichés, on peut très bien se contenter d'un ensemble $\Omega' = \{2, 3, 4 \ldots 12\}$.

---

1 Un exposé complet des fondements théoriques, comprenant en particulier le théorème de prolongement, dépasserait le cadre de ce livre. On se reportera à l'ouvrage de J. Neveu (1964).

Un ***événement*** est une assertion ou proposition logique relative au résultat de l'expérience (ex. : la somme des points est supérieure à 10). On dira qu'un événement est réalisé ou non suivant que la proposition est vraie ou fausse une fois l'expérience accomplie.

A la réalisation d'un événement on peut donc associer tous les résultats de l'épreuve correspondante ; ainsi la somme supérieure ou égale à 10 est l'ensemble de résultats suivants :

$$\{(4.6) ; (5.6) ; (6.6) ; (6.4) ; (6.5)\}$$

c'est-à-dire une partie de Ω. Désormais nous identifierons un événement à la partie de Ω pour laquelle cet événement est réalisé.

On appelle ***événement élémentaire*** une partie de Ω réduite à un seul élément.

1.1.2 Algèbre des événements

Réciproquement toute partie de Ω peut-elle être considérée comme un événement, ou du moins est-il utile qu'il en soit ainsi ? Afin de répondre à cette question nous allons supposer pour l'instant que l'ensemble des événements constitue une classe $\mathscr{C}$ de parties de Ω dont nous allons définir les propriétés en nous référant à des besoins usuels ; nous en profiterons pour introduire le vocabulaire probabiliste.

A tout événement A, on associe son contraire noté $\overline{A}$ tel que si A est réalisé alors $\overline{A}$ ne l'est pas, et réciproquement. $\overline{A}$ est donc représenté dans Ω par la partie complémentaire de A.

Il sera donc naturel d'exiger de $\mathscr{C}$ la propriété suivante : si $A \in \mathscr{C}$ alors $\overline{A} \in \mathscr{C}$.

Étant donné deux événements A, B on est conduit à s'intéresser à leur union A ou B ($A \cup B$) et à leur intersection (A et B ou $A \cap B$). Il faudra donc que si $A, B \in \mathscr{C}$, $A \cup B$ et $A \cap B \in \mathscr{C}$, et ceci d'une manière générale pour un nombre quelconque d'événements.

On définit également l'événement certain représenté par Ω tout entier et l'événement logiquement impossible (tel que avoir une somme de points égale à 13) représenté par l'ensemble vide $\varnothing$.

Nous pouvons maintenant définir la classe $\mathscr{C}$ par les trois axiomes :

- $\forall A \in \mathscr{C}, \overline{A} \in \mathscr{C}$;
- pour tout ensemble fini ou dénombrable $A_1, A_2, \ldots, A_n$ d'éléments de $\mathscr{C}$, $\bigcup_i A_i \in \mathscr{C}$;
- $\Omega \in \mathscr{C}$.

On peut montrer à titre d'exercice que ces axiomes impliquent que $\varnothing \in \mathscr{C}$ et que $\bigcap_i A_i \in \mathscr{C}$.

Les propriétés précédentes définissent ce que l'on appelle une σ-algèbre de Boole ou une tribu. $\mathscr{P}(\Omega)$ est une σ-algèbre particulière, la plus grosse, mais il n'est pas toujours utile ni souhaitable de l'utiliser.

On peut donc donner maintenant la définition d'un espace probabilisable :

DÉFINITION

On appelle espace probabilisable le couple $(\Omega ; \mathscr{C})$ *où* $\mathscr{C}$ *constitue une tribu de parties de* Ω.

Donnons encore quelques définitions utiles :

DÉFINITIONS

Événements incompatibles. *Deux événements A et B sont dits incompatibles si la réalisation de l'un exclut celle de l'autre, autrement dit si les parties A et B de* Ω *sont disjointes* $A \cap B = \varnothing$.

Système complet d'événements. $A_1, A_2, \ldots, A_n$ *forment un système complet d'événements si les parties* $A_1, \ldots, A_n$ *de* Ω *constituent une partition de* Ω :

$$\begin{cases} \forall i \neq j \qquad A_i \cap A_j = \varnothing \\ \cup A_i = \Omega \end{cases}$$

1.2 ESPACE PROBABILISÉ

1.2.1 L'axiomatique de Kolmogorov

A chaque événement on associe un nombre positif compris entre 0 et 1, sa probabilité. Afin d'éviter toute discussion de nature philosophique sur le hasard, la théorie moderne des probabilités repose sur l'axiomatique suivante :

DÉFINITIONS

On appelle ***probabilité sur*** $(\Omega, \mathscr{C})$ (***ou loi de probabilité***) *une application P de* $\mathscr{C}$ *dans* [0, 1] *telle que :*

- $P(\Omega) = 1$;
- *pour tout ensemble dénombrable d'événements incompatibles* $A_1, A_2, \ldots, A_n$, *on a* $P(\cup A_i) = \Sigma P(A_i)$.

On appelle ***espace probabilisé*** *le triplet* $(\Omega, \mathscr{C}, P)$.

Une loi de probabilité n'est donc rien d'autre qu'une mesure positive de masse totale 1 et la théorie des probabilités s'inscrit dans le cadre de la théorie de la mesure.

1.2.2 Propriétés élémentaires

Des axiomes on déduit immédiatement les propriétés suivantes :

Propriété 1 : $P(\varnothing) = 0$.

Propriété 2 : $P(\overline{A}) = 1 - P(A)$.

Propriété 3 : $P(A) \leq P(B) \qquad$ *si* $A \subset B$.

Propriété 4 : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Propriété 5 : $P(\cup A_i) \leq \sum_i P(A_i)$.

Propriété 6 : *Si* $A_i \downarrow \varnothing$, *alors* $\lim P(A_i) = 0$ (*continuité monotone séquentielle*).

Propriété 7 : *Théorème des probabilités totales : Soit* B_i *un système complet d'événements alors* $\forall A : P(A) = \sum_i P(A \cap B_i)$.

FORMULE DE POINCARÉ

Cette formule permet de calculer la probabilité de la réunion d'un nombre quelconque d'évenements ; elle se démontre par récurrence :

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i<j \leq 1}^{n} P(A_i \cap A_j)$$
$$+ \sum_{1 \leq i<j<k \leq 1}^{n} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} P(A_1 \cap A_2 \ldots \cap A_n)$$

Remarque : $P(A) = 0$ n'implique pas nécessairement $A = \varnothing$. Un événement de probabilité nulle n'est pas nécessairement impossible : soit $\Omega = [0, 1]$ muni de la loi de probabilité uniforme (c'est-à-dire de la mesure de Lebesgue) alors $P(\omega) = 0 \quad \forall \omega$.

De même $P(A) = 1$ n'implique pas que A soit l'événement certain : on parlera d'événement presque certain et dans le cas précédent d'événement presque impossible.

Les événements de probabilité nulle sont en réalité très communs, comme on le verra dans l'étude des variables aléatoires continues possédant une densité : tous les évènements $\{X = x\}$ sont de probabilité nulle mais aucun n'est impossible. La variable X prend une valeur précise une fois l'expérience réalisée. Cela est comparable au fait qu'un intervalle de longueur donnée l est formé d'une infinité de points de longueur nulle.

1.3 LOIS DE PROBABILITÉS CONDITIONNELLES, INDÉPENDANCE

Les concepts suivants sont purement probabilistes.

1.3.1 Introduction et définitions

Supposons que l'on s'intéresse à la réalisation d'un événement A, tout en sachant qu'un événement B est réalisé (fig. 1.1). Si A et B sont incompatibles la question est tranchée : A ne se réalisera pas, mais si $A \cap B \neq \varnothing$, il est possible que A se réalise ; cependant, l'univers des possibles n'est plus Ω tout entier, mais est restreint à B ; en fait, seule nous intéresse la réalisation de A à l'intérieur de B, c'est-à-dire $A \cap B$ par rapport à B.

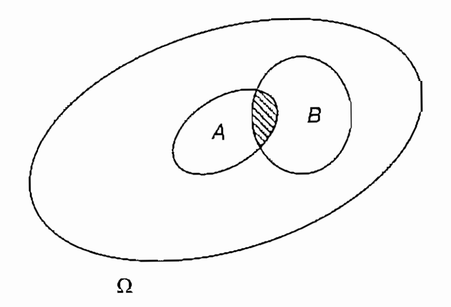


FIGURE 1.1

Ceci justifie la définition suivante :

DÉFINITION

Soit B un événement de probabilité non nulle. On appelle ***probabilité conditionnelle de A sachant B*** *(ou encore de A si B) le rapport noté P(A/B) :*

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Il faut s'assurer que le nom de probabilité est justifié. Vérifions les axiomes :

$$P(\Omega/B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$P\left(\bigcup_i A_i/B\right) = \frac{P\left[\left(\bigcup_i A_i\right) \cap B\right]}{P(B)} = \frac{P\left[\bigcup_i (A_i \cap B)\right]}{P(B)}$$

$$\sum_i \frac{P(A_i \cap B)}{P(B)} = \sum_i P(A_i/B) \qquad \text{c.q.f.d}$$

On peut donc munir $(\Omega, \mathscr{C})$ d'une nouvelle loi de probabilité, la loi de probabilité conditionnelle à B fixé et ceci pour tout B de probabilité non-nulle.

Il sera nécessaire d'étendre la notion de loi de probabilité conditionnelle lorsque B est de probabilité nulle (rappelons que la tribu $\mathscr{C}$ contient de tels événements) : cela sera fait au chapitre 3 dans certains cas particuliers.

Exemple : En fiabilité (ou en assurance sur la vie), on considère la fonction de survie $R(t)$ définie comme la probabilité qu'un individu vive au-delà d'une date t : $R(t) = P(X > t)$. Cette fonction définit une loi de probabilité sur $\mathbb{R}^+$ et :

$$P(t_1 \leq X < t_2) = R(t_1) - R(t_2)$$

La probabilité conditionnelle de défaillance (ou de décès) entre t_1 et t_2 sachant que l'individu a déjà fonctionné (ou vécu) jusqu'à t_1 est :

$$P(t_1 \leq X < t_2/X > t_1) = \frac{R(t_1) - R(t_2)}{R(t_1)}$$

Pour la loi de survie exponentielle $P(X > t) = \exp(-ct)$ on constate que cette probabilité conditionnelle vaut :

$$1 - \exp(-c\,(t_2 - t_1)) = P(X < t_2 - t_1)$$

il n'y a pas de vieillissement : la probabilité de fonctionner pendant $t_2 - t_1$ à partir de t_1 est la même qu'au démarrage. Ce modèle est couramment utilisé en électronique.

1.3.2 Indépendance

1.3.2.1 Indépendance de deux événements

DÉFINITION

A est indépendant de B si $P(A/B) = P(A)$.

Autrement dit, la connaissance de B ne change pas les « chances » de réalisation de A.

PROPRIÉTÉ

A indépendant de $B \Rightarrow B$ indépendant de A.

On parlera désormais d'événements indépendants sans autre précision.

En effet, si $P(A/B) = P(A)$, alors :

$$\frac{P(A \cap B)}{P(B)} = P(A)$$

et :

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = P(B)$$

On a démontré au passage l'importante formule :

$$\boxed{P(A \cap B) = P(A)P(B)}$$

si et seulement si A et B sont indépendants.

N.B. : La notion d'indépendance n'est pas une notion purement ensembliste comme l'incompatibilité : deux événements peuvent être indépendants pour une loi de probabilité P et pas pour une autre P'. On s'en convaincra en vérifiant qu'en général si A et B sont indépendants, ils ne le sont plus conditionnellement à un troisième événement C.

1.3.2.2 Indépendance deux à deux et indépendance mutuelle

Soient $A_1, A_2, \ldots, A_n$ des événements ; ils sont dits mutuellement indépendants si pour toute partie I de l'ensemble des indices allant de 1 à n on a :

$$P\left[\bigcap_I A_i\right] = \prod_I P(A_i)$$

Cette condition est beaucoup plus forte que l'indépendance deux à deux, qui ne lui est pas équivalente mais en est une simple conséquence.

Remarque : Dans les applications il est assez fréquent que l'on n'ait pas à démontrer l'indépendance de deux événements car celle-ci est une propriété de l'expérience aléatoire. Ainsi lorsqu'on procède à un tirage avec remise de n individus dans une population finie les événements relatifs aux différents tirages sont indépendants entre eux par construction.

1.3.3 Formules de Bayes

Elles ont pour but d'exprimer $P(A/B)$ en fonction de $P(B/A)$.

Première formule de Bayes :

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)}$$

Il suffit d'éliminer $P(A \cap B)$ entre $P(A/B) = \frac{P(A \cap B)}{P(B)}$ et $P(B/A) = \frac{P(A \cap B)}{P(A)}$.

Soit B_i un système complet d'événements. On peut écrire : $P(A \cap B_i) = P(A/B_i)P(B_i)$.

Le théorème des probabilités totales devient donc :

$$P(A) = \sum_i P(A/B_i)P(B_i)$$

On en déduit alors la ***deuxième formule de Bayes*** :

$$P(B_i/A) = \frac{P(A/B_i)P(B_i)}{\sum_k P(A/B_k)P(B_k)}$$

Exemple : Dans une usine trois machines M_1, M_2, M_3 fabriquent des boulons de même type. M_1 sort en moyenne 0.3 % de boulons défectueux, M_2 0.8 % et M_3 1 %. On mélange 1 000 boulons dans une caisse, 500 provenant de M_1, 350 de M_2 et 150 de M_3. On tire un boulon au hasard dans la caisse ; il est défectueux. Quelle est la probabilité qu'il ait été fabriqué par M_1 (ou M_2 ou M_3) ?

Lorsque l'on tire un boulon au hasard les probabilités dites *a priori* qu'il provienne de M_1, M_2 ou M_3 sont évidemment $P(M_1) = 0.50$, $P(M_2) = 0.35$, $P(M_3) = 0.15$.

Lorsque l'on sait qu'il est défectueux, événement noté D, il faut alors calculer les probabilités conditionnelles :

$$P(M_1/D),\ P(M_2/D),\ P(M_3/D)$$

Comme on connaît $P(D/M_1) = 0.003$, $P(D/M_2) = 0.008$ et $P(D/M_3) = 0.01$ la deuxième formule de Bayes permet d'écrire :

$$\begin{aligned} P(M_1/D) &= \frac{P(D/M_1)P(M_1)}{P(D/M_1)P(M_1) + P(D/M_2)P(M_2) + P(D/M_3)P(M_3)} \\ &= \frac{0.003 \times 0.5}{0.003 \times 0.5 + 0.008 \times 0.35 + 0.01 \times 0.15} \\ &\simeq 0.26 \end{aligned}$$

On trouverait de même $P(M_2/D) \simeq 0.48$ $\quad P(M_3/D) \simeq 0.26$.

Ce sont les probabilités *a posteriori*, sachant que le boulon est défectueux. On voit donc que la prise en compte d'une information (le boulon est défectueux) modifie les valeurs des probabilités de M_1, M_2 et M_3.

Le théorème de Bayes, simple conséquence des axiomes et de la définition de la probabilité conditionnelle, tient une place à part dans le calcul des probabilités en raison de son importance pratique considérable et des controverses auxquelles son application a donné lieu : il est à la base de toute une branche de la statistique appelée ***statistique bayésienne.***

Parmi les applications courantes citons : en diagnostic médical la révision des probabilités de telle ou telle affection après obtention des résultats d'examens de laboratoire, en matière financière la détermination du risque de faillite des entreprises après observations de certains ratios.

Le théorème de Bayes est souvent appelée théorème sur la « probabilité des causes » ce qui se conçoit aisément sur l'exemple précédent. Son application générale a donné lieu à de violentes critiques de la part des logiciens pour qui causalité et aléatoire sont antinomiques : il n'y a qu'une cause possible parmi des causes mutuellement exclusives et leur donner des probabilités n'aurait aucun sens.

Certains auteurs interprètent le fait que les formules de Bayes ont été publiées à titre posthume (en 1763) par la crainte du sacrilège : Thomas Bayes était en effet un ecclésiastique et l'application de sa formule à la recherche des causes ultimes d'un événement aurait pu conduire à probabiliser l'existence de Dieu. . .

1.4 RÉFLEXIONS SUR LE CONCEPT DE PROBABILITÉ

La théorie mathématique des probabilités ne dit pas quelle loi de probabilité mettre sur un ensemble Ω parmi toutes les lois possibles (et elles sont nombreuses. . .). Ce problème concerne ceux qui veulent appliquer le calcul des probabilités, et renvoie à la nature « physique », si l'on peut dire, du concept de probabilité qui formalise et quantifie le sentiment d'incertitude vis-à-vis d'un événement.

1.4.1 La conception objectiviste

Pour les tenants de ce point de vue, la probabilité d'un événement peut être déterminée de manière unique.

1.4.1.1 La vision classique

C'est celle qui est héritée des jeux de hasard. Ω est en général fini et des raisons de symétrie conduisent à donner à chaque événement élémentaire la même probabilité : ainsi le lancer d'un dé parfait conduit à un ensemble Ω à 6 éléments équiprobables.

Le calcul des probabilités n'est donc plus qu'une affaire de dénombrement, d'où la célèbre formule :

$$P(A) = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}$$

L'analyse combinatoire fournit alors les réponses aux cas classiques.

Cette approche ne s'étend pas aux cas où Ω n'est plus dénombrable (voir plus loin) et repose sur une conception idéalisée de l'expérience aléatoire : les symétries parfaites n'existent pas ; ainsi le dé parfait n'est qu'une vue de l'esprit et ses 6 faces ne sont pas en réalité

équiprobables en raison de la non homogénéité de la matière et surtout des gravures des numéros sur les faces.

1.4.1.2 Un paradoxe célèbre

Les limites de la vision classique apparaissent, nous semble-t-il, assez bien dans le célèbre paradoxe de Bertrand.

Considérons un triangle équilatéral et son cercle circonscrit. On tire une corde au hasard. Quelle est la probabilité que sa longueur soit supérieure à celle du côté du triangle ?

Reproduisons ici les commentaires de Renyi (1966) :

- **Première solution.** Comme la longueur de la corde est déterminée par la position de son milieu, le choix de la corde peut consister à marquer un point au hasard à l'intérieur du cercle. La probabilité pour que la corde soit plus longue que le côté du triangle équilatéral inscrit est alors évidemment égale à la probabilité pour que le milieu de la corde soit intérieur au cercle inscrit qui est de rayon moitié (*cf.* fig. 1.2).

Si l'on admet que la répartition de ce point est uniforme dans le cercle, on trouve pour la probabilité demandée :

$$\frac{\pi(r/2)^2}{\pi r^2} = \frac{1}{4}$$

- **Deuxième solution.** La longueur de la corde est déterminée par la distance de son milieu au centre du cercle. Par raison de symétrie nous pouvons considérer que le milieu de la corde est pris sur un rayon donné du cercle et supposer que la répartition de ce point sur le rayon est uniforme. La corde sera plus longue que le côté du triangle équilatéral inscrit si son milieu est à une distance du centre inférieur à $r/2$; la probabilité cherchée est alors $1/2$ (*cf.* fig. 1.3).

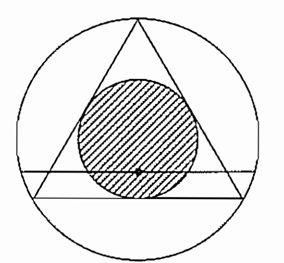


FIGURE 1.2

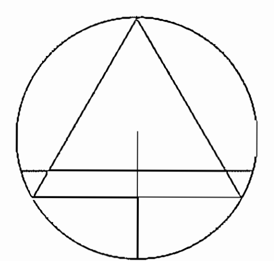


FIGURE 1.3

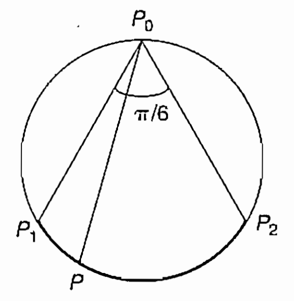


FIGURE 1.4

• Troisième solution. Par raison de symétrie nous pouvons supposer qu'on a fixé une des extrémités de la corde, soit P_0. L'autre sera choisie au hasard sur la circonférence. Si l'on admet que la probabilité pour que l'autre extrémité P tombe sur un arc donné de la circonférence est proportionnelle à la longueur de cet arc, la corde P_0P est plus grande que le côté du triangle équilatéral inscrit quand P se trouve sur l'arc P_1P_2 donc la longueur est le $1/3$ de celle de la circonférence (*cf.* fig. 1.4) ; la probabilité est alors $1/3$.

Il est clair que ces trois hypothèses de répartition, sont également réalisables. L'exemple parut paradoxal en son temps uniquement parce qu'on ne comprenait pas que des conditions expérimentales différentes pour le choix au hasard de la corde, dans les trois procédés décrits, conduisaient à des mesures-probabilités différentes sur la même algèbre d'événements.

1.4.1.3 La vision fréquentiste

Elle repose sur la loi des grands nombres (voir chapitre 2). Une seule expérience ne suffisant pas pour évaluer la probabilité d'un événement on va répéter un très grand nombre de fois l'expérience. Ainsi du lancer d'un dé : la probabilité d'observer la face 6 est la limite du rapport :

$$\frac{\text{Nombre de 6 obtenus}}{\text{Nombre d'essais}} = f$$

lorsque le nombre d'essais augmente indéfiniment. En effet la loi des grands nombres assure que f converge vers la probabilité p de l'événement.

Du point de vue pratique il est clair que la vision fréquentiste ne permet pas de trouver la probabilité d'un événement puisqu'un tel processus nécessitant une infinité d'observations est physiquement irréalisable : cela permet tout au plus de donner une définition de la probabilité comme limite d'une fréquence. Remarquons que dans la conception fréquentiste il est impossible de donner une valeur et même un sens à la probabilité d'un événement non répétable du genre « neigera-t-il le 25 octobre 2990 » ; ce qui limite le champ d'application du calcul des probabilités.

Cependant la critique la plus radicale du point de vue fréquentiste est la suivante : la définition de la probabilité repose sur la loi des grands nombres, or celle-ci est un théorème de probabilités qui suppose donc défini le concept de probabilité : il y a donc un cercle vicieux.

1.4.2 La conception subjectiviste

Le point de vue classique étant trop limité, le fréquentisme logiquement intenable, la probabilité d'un événement sujette à révision en fonction d'informations nouvelles (théorème de Bayes), l'existence même de probabilités objectives a été niée par beaucoup. C'est ainsi que le magistral Traité de Probabilités de de Finetti (1974) commence par l'affirmation en lettres capitales « *La Probabilité n'existe pas* » et continue par :

« *L'abandon de croyances superstitieuses sur l'existence du phlogistique, de l'éther, de l'espace et du temps absolu. . . ou des fées, a été une étape essentielle dans la pensée scientifique. La probabilité, considérée comme quelque chose ayant une existence objective est également une conception erronée et dangereuse, une tentative d'extérioriser ou de matérialiser nos véritables conceptions probabilistes !* »

1.4.2.1 Mesure d'incertitude

La probabilité objective d'un événement n'existe pas et n'est donc pas une grandeur mesurable analogue à la masse d'un corps, c'est simplement une ***mesure d'incertitude***, pouvant varier avec les circonstances et l'observateur, donc ***subjective***, la seule exigence étant qu'elle satisfasse aux axiomes du calcul des probabilités.

Les tenants de l'école subjectiviste proposent alors des méthodes permettant de passer d'une probabilité qualitative c'est-à-dire d'un simple pré-ordre sur les événements, à une mesure de probabilité.

Puisque la répétition n'est plus nécessaire on peut probabiliser des événements non répétables et étendre le domaine d'application du calcul des probabilités en particulier pour tout ce qui concerne les décisions économiques.

1.4.2.2 Le bayésianisme

Un pas de plus va être franchi par l'école bayésienne (ou plus exactement néo-bayésienne vu les deux siècles de décalage entre Bayes et ceux qui s'en réclament actuellement) qui va probabiliser tout ce qui est incertain et même des phénomènes non aléatoires.

Pour illustrer la théorie bayésienne modifions quelque peu l'exemple précédent de la fabrication des boulons : supposons qu'il n'y ait plus qu'une machine et que l'on cherche à estimer le pourcentage p de boulons défectueux produit en moyenne par la machine : si l'on admet qu'il n'y a que trois valeurs possibles p_1, p_2, p_3 respectivement égales à 0.3 %, 0.8 %, 1 % de probabilités *a priori* π_1, π_2, π_3 respectivement, la solution est inchangée et la valeur la plus probable *a posteriori* est 0.008 (si l'on tire un seul bouton défectueux). Supposons qu'on tire maintenant n boulons et que le nombre de boulons défectueux soit k, la probabilité que le pourcentage de défectueux produit par la machine soit p_2 est alors :

$$\frac{C_n^k p_2^k (1-p_2)^{n-k} \pi_2}{\sum_{i=1}^{3} C_n^k p_i^k (1-p_i)^{n-k} \pi_i}$$

On peut encore généraliser et supposer que p prenne toutes les valeurs possibles dans l'intervalle $[0, 1]$. Si l'on connaît la loi de probabilité de p sur $[0, 1]$ et qu'elle admet une densité $f(p)$ par rapport à la mesure de Lebesgue, la formule de Bayes s'écrit :

$$P(p/k) = \frac{C_n^k p^k (1-p)^{n-k} f(p)}{\int_0^1 C_n^k p^k (1-p)^{n-k} f(p)\, dp}$$

(voir chapitre 3).

A condition de connaître une distribution de probabilité *a priori* sur les valeurs de p, on peut donc en déduire les valeurs de p *a posteriori* les plus probables, donc estimer p.

On aura remarqué que p n'est pas aléatoire mais un paramètre fixe de valeur inconnue et que l'on a modélisé notre incertitude sur ses valeurs, par une mesure de probabilité. Mais

comment choisir cette mesure *a priori ?* on retombe sur la difficulté signalée plus haut et, si cette probabilité est subjective, quel statut scientifique donner à une grandeur qui peut varier d'un observateur à l'autre ? Telles sont les critiques usuelles faites par les objectivistes. De plus on a montré qu'un ordre de probabilités donné n'induisait pas nécessairement une mesure de probabilité unique P sur Ω, compatible avec la relation d'ordre. P n'existe pas forcément ou encore, si P existe, P n'est pas toujours unique.

Nous arrêterons là ces quelques remarques et sans prendre parti dans une querelle qui dure encore, rappelons que le modèle probabiliste a prouvé son efficacité dans de nombreuses applications mais que comme tout modèle ce n'est qu'une représentation simplificatrice de la réalité et que ses hypothèses doivent être mises à l'épreuve des faits.

Nous renvoyons le lecteur intéressé par la philosophie des probabilités aux travaux de de Finetti (1974), Matalon (1967), Matheron (1978) et Savage (1954), cités en références.

2 Variables aléatoires

Dans ce chapitre, on étudiera uniquement les variables aléatoires réelles. Les variables qualitatives ou ordinales (à valeurs dans un ensemble quelconque ou muni d'une structure d'ordre) ne feront pas l'objet d'une étude théorique ; on les trouvera évoquées dans les chapitres consacrés à la statistique.

2.1 LOI DE PROBABILITÉ ET MOMENTS D'UNE VARIABLE ALÉATOIRE RÉELLE

2.1.1 Définition et fonction de répartition

2.1.1.1 Généralités

Le concept de variable aléatoire formalise la notion de grandeur variant selon le résultat d'une expérience aléatoire.

Considérons le lancer de deux dés parfaitement équilibrés : cette expérience se traduit par l'ensemble Ω de tous les couples de chiffres de 1 à 6 :

$$\Omega = \{(1, 1) ; (1, 2) ; \ldots ; (6, 6)\}$$

muni de la loi de probabilité P telle que $P(\omega) = \dfrac{1}{36}, \forall \omega \in \Omega$.

Intéressons-nous à la somme des points marqués par les deux dés. On définit ainsi une application S de Ω dans l'ensemble $E = \{2, 3, \ldots, 12\}$ (fig. 2.1).

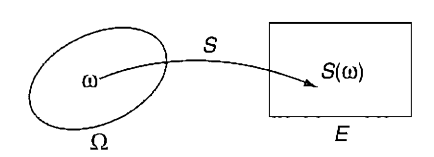


FIGURE 2.1

Pour obtenir la probabilité d'une valeur quelconque de S, il suffit de dénombrer les ω qui réalisent cette valeur. Ainsi :

$$P(S = 5) = P(\{(1,4)(2,3)(3,2)(4,1)\}) = \frac{4}{36}$$

et généralement $P(S = s) = P(\{S^{-1}(s)\})$.

On voit que, pour définir la loi de probabilité sur S, on transporte la loi de probabilité de Ω sur E par l'application S.

Si X est une application d'un ensemble probabilisé $(\Omega, \mathscr{C}, P)$ dans E, il faut donc que E soit probabilisable, c'est-à-dire muni d'un tribu $\mathscr{T}$ et que l'image réciproque de tout élément de $\mathscr{T}$ soit un événement, c'est-à-dire un élément de $\mathscr{C}$. On reconnaît ici la définition mathématique de la mesurabilité d'une fonction.

Une variable aléatoire X est donc une application mesurable de $(\Omega, \mathscr{C}, P)$ dans $(E, \mathscr{T})$.

Lorsque $E = \mathbb{R}$, on utilise comme tribu la σ-algèbre engendrée par les intervalles de $\mathbb{R}$; c'est la plus petite σ-algèbre (autrement dit l'intersection de toutes les σ-algèbres) contenant les intervalles. Cette tribu est appelée tribu borélienne et est notée $\mathscr{B}$.

DÉFINITION 1

Une variable aléatoire réelle est une application mesurable de $(\Omega, \mathscr{C}, P)$ dans $\mathbb{R}$ muni de sa tribu borélienne $(\mathbb{R}, \mathscr{B})$.

Pour tout borélien B, on définit $P_X(B)$ par :

$$\begin{aligned} P_X(B) &= P(\{\omega | X(\omega) \in B\}) \\ &= P([X^{-1}(B)]) \end{aligned}$$

ceci définit une probabilité sur $(\mathbb{R}, \mathscr{B})$ d'où la :

DÉFINITION 2

On appelle loi de probabilité de X la mesure image de P par X et on la note P_X.

Pour une variable discrète, c'est-à-dire une variable ne pouvant prendre qu'un nombre fini (ou dénombrable) de valeurs $x_1, x_2, \ldots, x_n$, la loi P_X est constituée de masses ponctuelles. P_X peut alors être représentée par un diagramme en bâtons.

Ainsi, pour l'exemple du lancer de deux dés, on a la figure 2.2.

2.1.1.2 Fonction de répartition

La fonction de répartition d'une variable aléatoire X est l'application F de $\mathbb{R}$ dans $[0, 1]$ définie par :

$$\boxed{F(x) = P(X < x)}$$

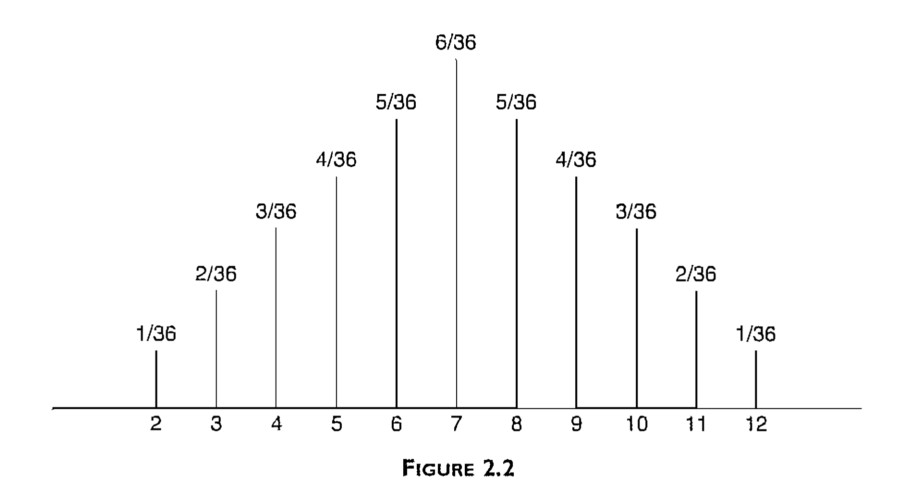


FIGURE 2.2

PROPRIÉTÉS (sans démonstration)

> F est une fonction monotone croissante continue à gauche. En tant que fonction monotone, elle admet un nombre de points de discontinuité au plus dénombrable. Réciproquement, toute fonction monotone croissante continue à gauche telle que $F(-\infty) = 0$ et $F(+\infty) = 1$ définit une loi de probabilité unique sur $\mathbb{R}$.

Un exemple de fonction de répartition correspondant à une variable discrète (celle de S définie précédemment) est donné par la figure 2.3.

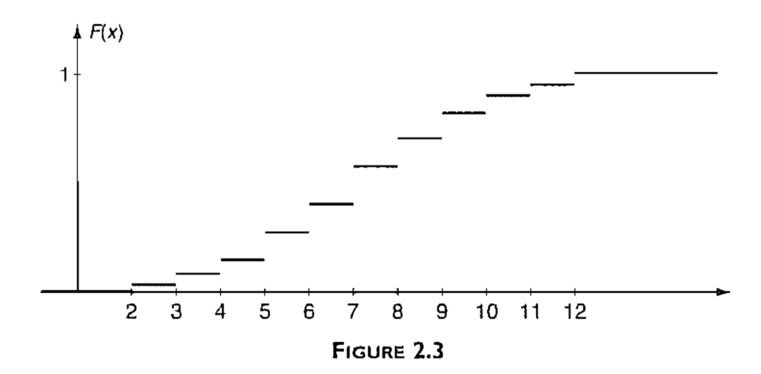


FIGURE 2.3

La figure 2.4 est un exemple de fonction de répartition correspondant à une variable continue (voir plus loin).

L'importance pratique de la fonction de répartition est qu'elle permet de calculer la probabilité de tout intervalle de $\mathbb{R}$:

$$\boxed{P(a \leq X < b) = F(b) - F(a)}$$

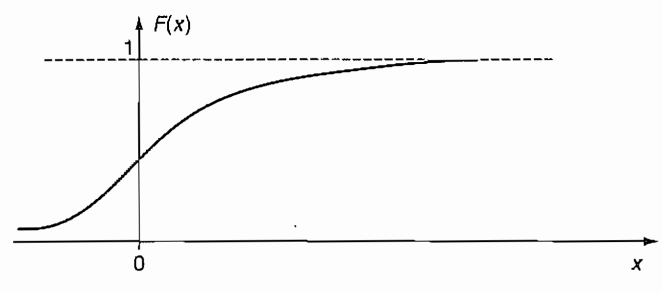


Figure 2.4

2.1.1.3 Variables continues

La notion de variable continue, ou plus exactement absolument continue, se confond avec celle de variable admettant une densité de probabilité.

Définition

Une loi de probabilité P_X admet une densité f si, pour tout intervalle I de $\mathbb{R}$, on a :

$$P_X(I) = \int_I f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \mathbb{1}_I(x) f(x)\,\mathrm{d}x$$

($\mathbb{1}_I$ est la fonction indicatrice de I).

F est alors dérivable et admet f pour dérivée. On a donc :

$$\boxed{P(a < X < b) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)} \qquad \text{(fig. 2.5)}$$

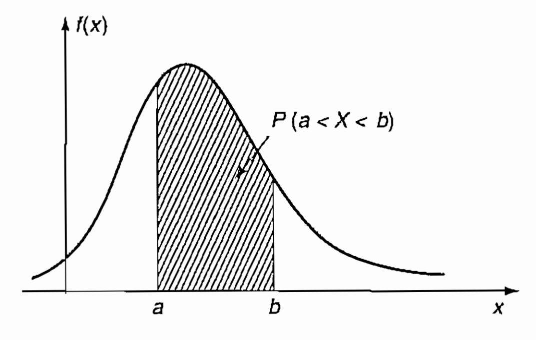


Figure 2.5

Une densité f est donc une fonction positive d'intégrale égale à 1 :

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$$

On remarque que pour une variable à densité :

$$P(X = x) = 0 \qquad \forall x$$

et on peut écrire : $$\boxed{P(x < X < x + \mathrm{d}x) = f(x)\,\mathrm{d}x}$$

■ ***Exemple :*** La variable X, dont la loi est définie par $P(X > x) = \exp(-\lambda x)$ pour tout x positif, admet pour densité :

$$f(x) = \lambda \exp(-\lambda x) \quad \text{si } x \geq 0$$
$$f(x) = 0 \quad \text{si } x < 0 \qquad \text{(fig. 2.6)}$$

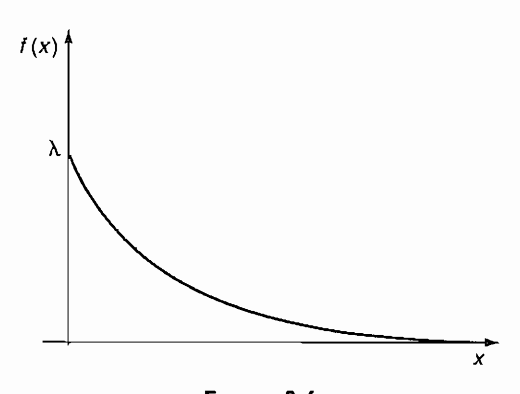


FIGURE 2.6

Elle est utilisée couramment pour représenter la durée de vie de phénomènes sans vieillissement (comme les composants électroniques). ■

2.1.1.4 *Taux instantané de défaillance*

Si X est une variable continue positive représentant une durée, on définit la fonction suivante :

$$h(x) = \frac{f(x)}{1 - F(x)}$$

appelées selon les domaines d'application : « taux instantané de défaillance », « fonction de hasard » ou encore « quotient de mortalités ». Pour une durée de vie X, $h(x)$ s'interprète comme la probabilité de décès immédiatement après x, sachant que l'on a vécu jusqu'à x.

En effet, pour dx infiniment petit :

$$P(x < X < x + \mathrm{d}x / X > x) = \frac{f(x)\,\mathrm{d}x}{1 - F(x)} = h(x)\,\mathrm{d}x.$$

$1 - F(x)$ est appelée fonction de survie.

$h(x)$ caractérise la loi de X car on peut retrouver $F(x)$ à partir de $h(x)$:

$$h(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\ln(1 - F(x))$$

$$F(x) = 1 - \exp(-\int_0^x h(t)\mathrm{d}t)$$

Une fonction $h(x)$ croissante est caractéristique d'un phénomène de vieillissement.

Si $h(x) = c$, il y a absence de vieillissement, le décès est dû à des causes aléatoires externes : X suit alors la loi exponentielle $F(x) = 1 - \exp(-cx)$, qui sera étudiée plus loin.

2.1.2 Loi d'une fonction d'une variable aléatoire $Y = \varphi(X)$

On supposera X continue avec une densité f et une fonction de répartition F. φ sera supposé dérivable. On recherche g et G densité et fonction de répartition de Y.

2.1.2.1 φ *bijective*

φ est donc monotone. Si φ est croissante, on a $F(x) = G(\varphi(x))$ car $X < x \Leftrightarrow Y < \varphi(x)$ d'où :

$$\boxed{G(y) = F(\varphi^{-1}(y))} \qquad \text{(fig. 2.7a)}$$

En dérivant : $f(x) = g(\varphi(x))\varphi'(x)$ soit :

$$g(y) = \frac{f(x)}{\varphi'(x)}$$

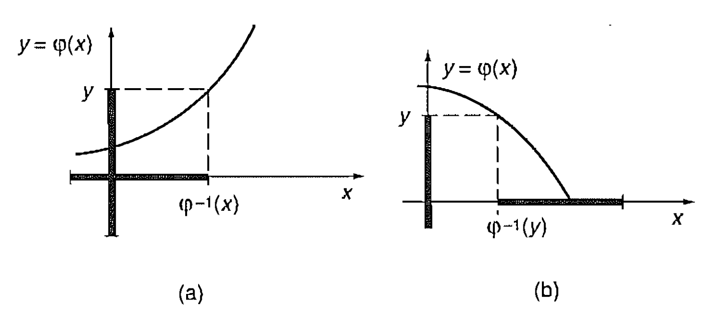


FIGURE 2.7a FIGURE 2.7b

ou encore :

$$g(y) = \frac{f[\varphi^{-1}(y)]}{\varphi'[\varphi^{-1}(y)]}$$

Si φ est décroissante $X < x \Leftrightarrow Y > \varphi(x)$, d'où :

$$\boxed{G(y) = 1 - F(\varphi^{-1}(y))} \qquad \text{(fig. 2.7b)}$$

et en dérivant : $$g(y) = -\frac{f(x)}{\varphi'(x)}$$

Puisque φ est décroissante, $\varphi' < 0$, et on a la formule générale pour une application bijective φ quelconque :

$$g(y) = \frac{f(x)}{|\varphi'(x)|}$$

$$\boxed{g(y) = \frac{f[\varphi^{-1}(y)]}{|\varphi'[\varphi^{-1}(y)]|}}$$

Exemple : $Y = \exp(X)$ et $X = \ln Y$

$$g(y) = \frac{f(x)}{\exp(x)} = \frac{f(\ln y)}{y}$$

2.1.2.2 *φ quelconque*

Le principe consiste toujours à identifier la fonction de répartition $G(y)$ en recherchant l'antécédent pour X de l'événement $Y < y = \varphi(x)$.

Par exemple, si $Y = X^2$ avec X défini sur $\mathbb{R}$: $P(Y < y) = P(-\sqrt{y} < X < +\sqrt{y})$:

$$\boxed{G(y) = F(\sqrt{y}) - F(-\sqrt{y})}$$

$$g(y) = f(\sqrt{y})\frac{1}{2\sqrt{y}} + f(-\sqrt{y})\frac{1}{2\sqrt{y}}$$

$$\boxed{g(y) = \frac{1}{2\sqrt{y}}(f(\sqrt{y}) + f(-\sqrt{y}))}$$

en particulier $g(y) = \dfrac{f(\sqrt{y})}{\sqrt{y}}$ si f est une fonction paire.

2.1.3 Indépendance de deux variables aléatoires

Soient X et Y deux variables aléatoires réelles définies sur le même espace probabilisé. Le couple (X, Y) est donc une application mesurable de $(\Omega, \mathscr{C}, P)$ dans $\mathbb{R}^2$ muni de sa tribu borélienne.

DÉFINITION

X et Y sont indépendantes si, pour tout couple de boréliens B_i et B_j, on a :

$$P((X \in B_i) \cap (Y \in B_j)) = P(X \in B_i)P(Y \in B_j)$$

En d'autres termes, la loi de probabilité P_{XY} du couple (X, Y) n'est autre que la loi produit que l'on note :

$$P_{XY} = P_X \otimes P_Y$$

COROLLAIRE

X et Y sont indépendantes si et seulement si la fonction de répartition du couple (X, Y) définie par $H(x, y) = P(X < x \cap Y < y)$ est égale au produit des fonctions de répartition respectives de X et de Y, appelées fonctions de répartition marginales :

$$\boxed{H(x, y) = F(x)\,G(y)}$$

Si X et Y admettent des densités $f(x)$ et $g(y)$, alors le couple (X, Y) admet pour densité $f(x)g(y)$. Dans ce cas, la réciproque est également vraie.

2.1.4 Moments d'une variable aléatoire

Une loi de probabilité peut être caractérisée par certaines valeurs typiques associées aux notions de valeur centrale, de dispersion et de forme de la distribution.

2.1.4.1 L'espérance mathématique

Pour une variable discrète, on définit l'espérance $E(X)$ par la formule :

$$E(X) = \sum_i x_i P(X = x_i)$$

(si cette expression a un sens). $E(X)$ est la moyenne arithmétique des différentes valeurs de X pondérées par leurs probabilités.

Pour une variable continue admettant une densité, $E(X)$ est la valeur, si l'intégrale converge, de $\int_{\mathbb{R}} x f(x)\,\mathrm{d}x$.

Ces deux expressions ne sont en fait que des cas particuliers de la définition générale suivante :

DÉFINITION

X étant une variable aléatoire réelle définie sur $(\Omega, \mathcal{C}, P)$, l'espérance mathématique de X est, si elle existe, l'intégrale de X par rapport à la mesure P :

$$E(X) = \int_{\Omega} X\,\mathrm{d}P$$

D'après le théorème de la mesure image, on a :

$$\boxed{E(X) = \int_{\mathbb{R}} x\,\mathrm{d}P_X(x)}$$

d'où, en particulier si P_X est absolument continue par rapport à la mesure de Lebesgue de $\mathbb{R}$, il existe une densité $f(x)$: $dP_X(x) = f(x)\,dx$ et alors on retrouve :

$$E(X) = \int_{\mathbb{R}} x f(x)\,dx$$

Il faut prendre garde au fait que l'espérance mathématique n'existe pas toujours. Ainsi, la variable X ayant pour densité sur $\mathbb{R}$:

$$f(x) = \frac{1}{\pi(1 + x^2)} \quad \text{(loi de Cauchy)}$$

n'a pas d'espérance car l'intégrale $\displaystyle\int_{-\infty}^{+\infty} \frac{x}{\pi(1+x^2)}\,dx$ diverge.

Les propriétés élémentaires de l'espérance mathématique sont celles des intégrales et se déduisent de la linéarité. Si a est une constante :

$$\boxed{\begin{aligned} E(a) &= a \\ E(aX) &= aE(X) \\ E(X+a) &= E(X) + a \end{aligned}}$$

La plus importante propriété est l'additivité : l'espérance d'une somme de variables aléatoires (qu'elles soient ou non indépendantes) est égale à la somme de leurs espérances :

$$\boxed{E(X_1 + X_2) = E(X_1) + E(X_2)}$$

A. Espérance d'une fonction φ(X) d'une variable aléatoire

Par définition, $E[\varphi(X)] = \displaystyle\int_\Omega (\varphi \circ X)\,dP$ si cette expression a un sens.

En utilisant à nouveau le théorème de la mesure image, on a :

$$\boxed{E(\varphi(X)) = \int_{\mathbb{R}} \varphi(x)\,dP_X(x)}$$

Ce résultat très important est d'un emploi courant et permet de calculer l'espérance d'une variable $\varphi(X)$ sans avoir à déterminer la loi de $\varphi \circ X$.

B. Inégalité de Jensen

Si φ est une fonction convexe, on peut montrer, si les espérances existent, que :

$$\boxed{E(\varphi(X)) \geq \varphi(E(X))}$$

On en déduit en particulier :

$$E(|X|) \geq |E(X)|$$
$$E(X^2) \geq (E(X))^2$$
$$E(\exp(X)) \geq \exp(E(X))$$

C. Espérance d'un produit

Si X et Y sont deux variables aléatoires de loi conjointe P_{XY}, on a, si l'expression a un sens :

$$E(XY) = \int_{\mathbb{R}^2} xy \, \mathrm{d}P_{XY}(x, y)$$

Lorsque X et Y sont indépendants, $\mathrm{d}P_{XY}(x, y) = \mathrm{d}P_X(x) \otimes \mathrm{d}P_Y(y)$ et l'intégrale double se factorise :

$$E(XY) = \int_{\mathbb{R}} x \, \mathrm{d}P_X(x) \int_{\mathbb{R}} y \, \mathrm{d}P_Y(y)$$

d'où : $$\boxed{X \text{ et } Y \text{ indépendants} \Longrightarrow E(XY) = E(X)E(Y)}$$

Attention : La réciproque est fausse et $E(X)E(Y) = E(XY)$ n'entraîne pas en général l'indépendance de X et Y.

D. Une interprétation statistique

Reprenons l'exemple du lancer de deux dés. Par raison de symétrie, $E(S) = 7$. Supposons qu'on lance n fois les deux dés et que les réalisations successives de S soient $s_1, s_2, \ldots, s_n$. Formons la moyenne $\bar{s} = \dfrac{1}{n}\sum s_i$ de ces résultats.

On montre alors que si $n \longrightarrow \infty$, $\bar{s} \longrightarrow 7$ en un sens qui sera précisé plus tard (loi des grands nombres, voir paragr. 2.7 et chapitre 12).

E. Espérance et fonction de répartition

Sous réserve de convergence de l'intégrale, on a pour une variable ***positive*** le résultat suivant :

$$E(X) = \int_0^{\infty} (1 - F(x)) \, \mathrm{d}x$$

En effet, en intégrant par parties : $\int_0^{\infty} (1 - F(x)) \, \mathrm{d}x = [(1 - F(x))x]_0^{\infty} + \int_0^{\infty} x f(x) \, \mathrm{d}x$, et le crochet est nul si l'intégrale converge.

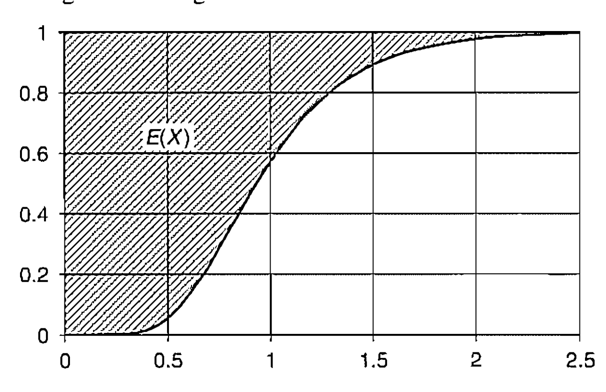


FIGURE 2.8

L'espérance d'une variable positive s'interprète donc comme l'aire située entre l'horizontale $y = 1$ et la fonction de répartition. La figure 2.8 correspond à la fonction de répartition d'une loi log-normale d'espérance 1 et d'écart-type 0.4.

2.1.4.2 La variance

On appelle variance de X notée $V(X)$ ou σ^2 la quantité définie par :

$$\boxed{\sigma^2 = E((X - m)^2) = \int_{\mathbb{R}} (x - m)^2 \, dP_X(x)}$$

où $m = E(X)$.

σ s'appelle l'***écart-type*** de X.

La variance est donc le moment centré d'ordre 2 de la distribution et est une mesure de la dispersion de X autour de m.

- Propriétés de la variance

Comme $E((X - a)^2) = V(X) + (E(X) - a)^2$ (formule de König-Huyghens) on en déduit que $V(X)$ est la valeur minimale de $E((X - a)^2)$ quand a varie.

On en déduit la formule classique

$$V(X) = E(X^2) - (E(X))^2$$

Par ailleurs :

$$V(X-a) = V(X)$$

$$V(aX) = a^2V(X) \quad \text{et} \quad \sigma(aX) = |a|\,\sigma(X)$$

$$V(X) = 0 \Leftrightarrow X = a \text{ (presque sûrement)}$$

L'espérance et l'écart-type sont reliés par *l'inégalité de* **Bienaymé-Tchebyshev** :

$$\boxed{P(|X - E(X)| > k\sigma) \leq \frac{1}{k^2}}$$

Démonstration

$$\sigma^2 = \int_{\mathbb{R}} (x - m)^2 \, dP_X(x) > \int_{|X-m|>k\sigma} (x - m)^2 \, dP_X(x)$$

car on restreint le domaine d'intégration d'une fonction positive. En minorant $(x - m)^2$ par $k^2\sigma^2$, on a :

$$\int_{|X-m|>k\sigma} (x - m)^2 \, dP_X(x) > k^2\sigma^2 \int_{|X-m|>k\sigma} dP_X(x)$$

Cette dernière intégrale vaut $P(|X - m| > k\sigma)$, ce qui établit la propriété.

Cette inégalité, dont l'intérêt théorique vient de ce qu'elle est valable quelle que soit la loi de X, n'a que peu d'applications pratiques, car la majoration qu'elle fournit est la plupart du temps excessive. Ainsi pour une loi normale, $P(|X - E(X)| > 2\sigma) \simeq 0.05$ alors que l'inégalité de Bienaymé-Tchebyshev donne 0.25 comme majorant. Remarquons, de plus, que l'inégalité est inutilisable pour $k \leq 1$.

- Variance d'une somme de variables aléatoires

$$\begin{aligned} V(X+Y) &= E[(X+Y)^2] - (E(X)+E(Y))^2 \\ &= E(X^2) + E(Y^2) + 2E(XY) - E(X)^2 - E(Y)^2 - 2E(X)E(Y) \\ &= V(X) + V(Y) + 2(E(XY) - E(X)E(Y)) \end{aligned}$$

On appelle covariance de X et Y la quantité :

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) = E((X - E(X))(Y - E(Y)))$$

donc :

$$\boxed{V(X+Y) = V(X) + V(Y) + 2\,\text{cov}(X, Y)}$$

En particulier : X et Y indépendantes $\Rightarrow V(X+Y) = V(X) + V(Y)$

mais la réciproque est ici encore inexacte en général.

- Variance d'un produit de deux variables indépendantes

Un calcul élémentaire montre que :

$$\boxed{V(XY) = V(X)V(Y) + V(X)(E(Y))^2 + V(Y)(E(X))^2}$$

- Approximations de l'espérance et de la variance d'une fonction $\varphi(X)$

Un développement limité à l'ordre 2 au voisinage de l'espérance m de X donne :

$$\varphi(x) - \varphi(m) \simeq (x-m)\varphi'(m) + \frac{(x-m)^2}{2}\varphi''(m)$$

En prenant l'espérance :

$$E(\varphi(X)) - \varphi(m) \simeq E\left(\frac{(X-m)^2}{2}\right)\varphi''(m)$$

soit :

$$\boxed{E(\varphi(X)) \simeq \varphi(m) + \frac{1}{2}V(X)\varphi''(m)}$$

En élevant au carré $\varphi(X) - \varphi(m)$ et en prenant l'espérance, on trouve également [Lejeune, 2004] :

$$\boxed{V(\varphi(X)) \simeq (\varphi'(m))^2 V(X)}$$

2.1.4.3 Autres moments

On définit, si ils existent, les moments centrés d'ordre k :

$$\boxed{\mu_k = E[(X - m)^k]}$$

On a évidemment $\mu_1 = 0$ et $\mu_2 = V(X)$. Si la distribution de la variable aléatoire est symétrique, on a $\mu_{2k+1} = 0 \quad \forall k$.

Les moments μ_3 et μ_4 sont utilisés pour caractériser la forme de distribution.

Pour obtenir des quantités sans dimension, on utilise les coefficients d'asymétrie et d'aplatissement γ_1 et γ_2 (en anglais *skewness* et *kurtosis*) :

$$\boxed{\gamma_1 = \frac{\mu_3}{\sigma^3}} \qquad \boxed{\gamma_1 = \frac{\mu_4}{\sigma^4}}$$

La figure 2.9 donne quelques allures typiques de courbes de densité correspondant à certaines valeurs de γ_1 et γ_2.

On remarquera que γ_2 est toujours supérieur à 1 car l'inégalité classique entre moyennes d'ordre p entraîne $(\mu_4)^{1/4} > (\mu_2)^{1/2} \Rightarrow \mu_4 > (\mu_2)^2$.

De plus, on a toujours $\gamma_2 \geq 1 + (\gamma_1)^2$.

Plus que l'aplatissement, le coefficient γ_2 mesure l'importance des « queues » de distribution.

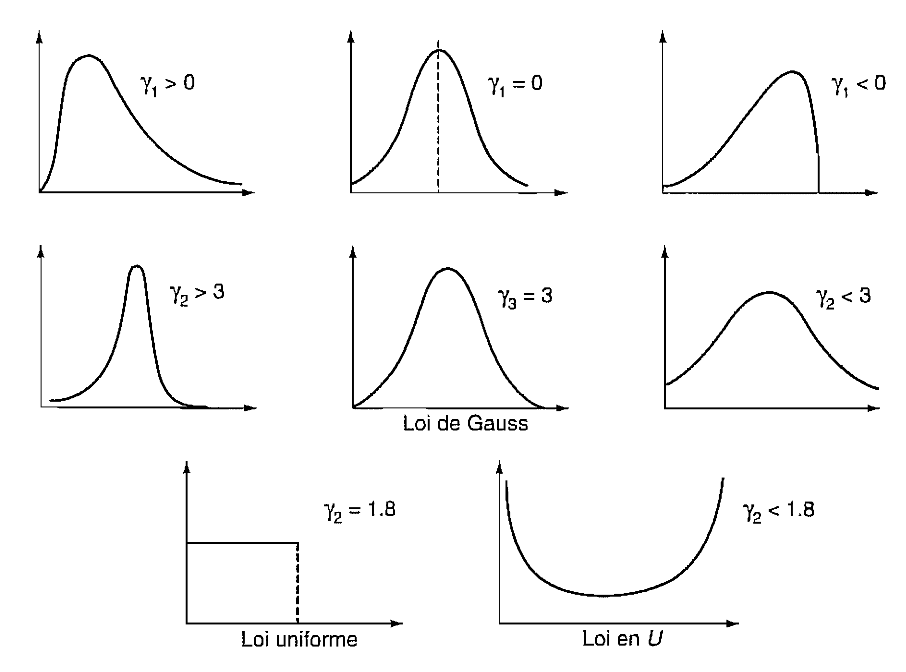


FIGURE 2.9

Inégalité de Markov : En utilisant la même méthode que pour l'inégalité de Bienaymé-Tchebyshev, on montre que :

$$P(|X|) > \varepsilon) \leq \frac{E(X^k)}{\varepsilon^k}$$

2.1.4.4 Ordres stochastiques

Les concepts de dominance stochastique sont utilisés dans différents domaines, en particulier en fiabilité pour comparer des fonctions de survie, et en théorie de la décision pour comparer des risques.

A. Dominance stochastique d'ordre 1

On dit que X domine stochastiquement Y si la fonction de survie de X est supérieure à celle de Y :

$$P(X > c) \geq P(Y > c) \text{ pour tout } c$$

ce qui revient à dire que la fonction de répartition de X est toujours inférieure à celle de Y.

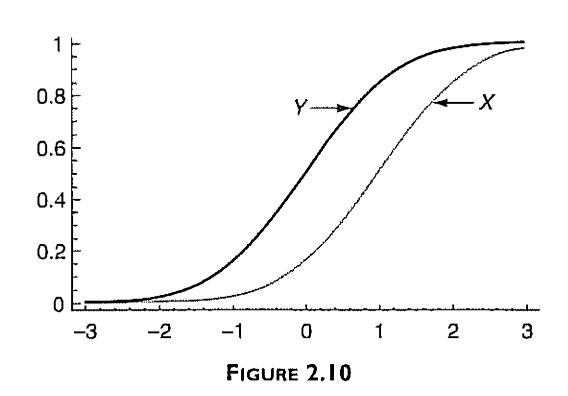


FIGURE 2.10

THÉORÈME (ADMIS)

Pour que X domine stochastiquement Y, il faut et il suffit que $E(f(X)) \geq E(f(Y))$ pour toute fonction f croissante.

On en déduit que la dominance stochastique de X sur Y entraine $E(X) \geq E(Y)$.

On peut montrer (exercice à faire . . .) la propriété suivante : si la fonction de hasard (ou taux de défaillance) de X est partout inférieure à celle de Y, alors X domine stochastiquement Y. C'est par exemple le cas de la durée de vie des femmes en France qui domine celle des hommes : non seulement l'espérance de vie des femmes est plus élevée que celle des hommes, mais également la probabilité de survie à tout âge.

B. Dominance stochastique d'ordre 2

La dominance d'ordre 1 implique que les fonctions de répartition de X et Y ne peuvent se croiser. Une forme plus faible de dominance, qui autorise les croisements, est définie comme suit :

DÉFINITION

X domine stochastiquement Y à l'ordre 2 si leurs fonctions de répartition sont telles que :

$$\int_{-\infty}^{c} F(x)\,\mathrm{d}x \leq \int_{-\infty}^{c} G(x)\,\mathrm{d}x \quad \textit{pour tout } c.$$

L'inégalité porte cette fois sur les intégrales des fonctions de répartition. La dominance stochastique d'ordre 1 entraîne celle d'ordre 2.

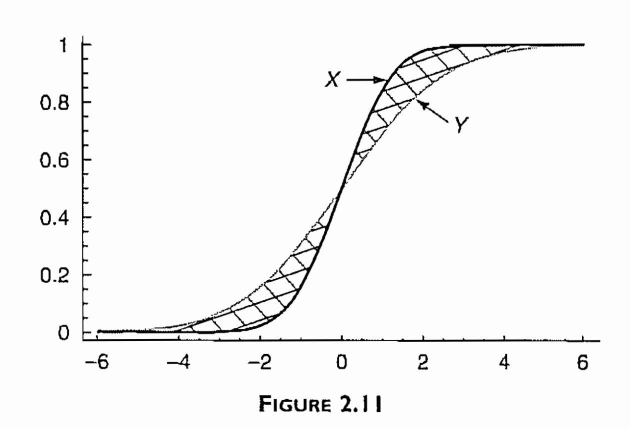


FIGURE 2.11

Cette forme de dominance est utilisée en théorie du risque pour des variables positives représentant des gains aléatoires. Supposons de plus que X et Y ont même espérance : alors les aires hachurées sur la figure précédente sont égales. On voit intuitivement que la répartition de X est moins dispersée que celle de Y. Un individu qui a de l'aversion pour le risque préfèrera donc X à Y. La dominance stochastique d'ordre 2 implique $V(X) < V(Y)$ mais est plus générale (la réciproque est fausse).

On montre que si X domine Y, Y a la même distribution que $X + \varepsilon$ où ε est une variable telle que $E(\varepsilon/X) = 0$. Intuitivement, Y est « plus aléatoire » que X.

Le théorème du paragraphe précédent est alors modifié comme suit [Rothschild et Stiglitz, 1970] :

THÉORÈME

Pour que X domine stochastiquement Y à l'ordre 2, il faut et il suffit que $E(f(X)) \geq E(f(Y))$ pour toute fonction f croissante concave.

2.2 LOIS DE PROBABILITÉ DISCRÈTES D'USAGE COURANT

2.2.1 Loi discrète uniforme

$$X = \{1, 2, 3, \ldots, n\}$$

$$P(X = 1) = P(X = 2) = \cdots = P(X = n) \qquad \text{(fig. 2.12)}$$

$$P(X = k) = \frac{1}{n}$$

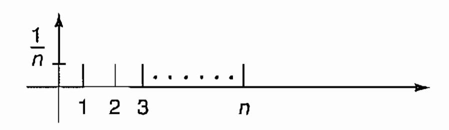


FIGURE 2.12

$$\boxed{E(X) = \frac{n+1}{2}} \qquad \text{par symétrie}$$

$$E(X) = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{n+1}{2}$$

$$E(X^2) = \frac{1}{n}(1 + 4 + 9 + \cdots + n^2)$$

$$E(X^2) = \frac{1}{n}\frac{n(n+1)(2n+1)}{6}$$

d'où :

$$V(X) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$V(X) = \frac{n+1}{12}(4n + 2 - 3(n+1))$$

soit :

$$\boxed{V(X) = \frac{n^2 - 1}{12}}$$

2.2.2 Loi de Bernoulli de paramètre p

C'est la loi d'une variable X ne pouvant prendre que les deux valeurs 1 ou 0 avec les probabilités p et $1 - p$; X est la fonction indicatrice d'un événement A de probabilité p :

$$\boxed{E(X) = p}$$

Comme $X^2 = X$, $E(X^2) = p$, d'où :

$$\boxed{V(X) = p(1 - p)}$$

2.2.3 Loi binomiale $\mathcal{B}(n\,;p)$

A. Principe

Supposons que l'on repète n fois dans des conditions identiques une expérience aléatoire, dont l'issue se traduit par l'apparition ou la non-apparition d'un événement A de probabilité p, le résultat de chaque expérience étant indépendant des résultats précédents. Soit X le nombre d'apparitions de l'événement A parmi ces n expériences ($0 \leq X \leq n$). On dit alors que X suit une loi binomiale de paramètres n et p notée $\mathcal{B}(n\,;p)$. Comme à chaque expérience numérotée $i (i = 1, 2, \ldots, n)$, on peut associer une variable de Bernoulli X_i de paramètre p, on a : $X = \sum_{i=1}^{n} X_i$ d'où la deuxième définition de la loi binomiale : X suit une loi binomiale $\mathcal{B}(n\,;p)$ si X est une somme de n variables de Bernoulli indépendantes et de même paramètre p.

De cette définition, découlent l'espérance et la variance de X.

$E(X) = \Sigma E(X_i)$, donc : $\boxed{E(X) = np}$ $V(X) = \Sigma V(X_i)$ car les X_i sont indépendants ; donc :

$$\boxed{V(X) = np(1-p)}$$

B. Loi de probabilité

Afin de chercher l'expression de $P(X = k)$, remarquons que toutes les configurations, telles que k variables X_i prennent la valeur 1 et $n - k$ la valeur 0, sont équiprobables et qu'il y a C_n^k configurations de cette sorte (nombre de manières de choisir k X_i parmi n).

D'autre part :

$$P(X_1 = x_1 \cap \ldots \cap X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$
$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

car les X_i sont indépendants :

$$P(X_1 = x_1 \cap X_2 = x_2, \ldots, \cap X_n = x_n) = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$$

Comme $\Sigma x_i = k$, on trouve :

$$\boxed{P(X = k) = C_n^k p^k (1-p)^{n-k}}$$

Cette formule justifie le nom de la loi binomiale car les $P(X = k)$ sont les termes du développement de $(p + (1-p))^n$ selon la formule du binôme de Newton (on vérifie au passage que $\sum_{k=0}^{k=n} P(X = k) = 1$).

La figure 2.13 représente quelques diagrammes en bâtons correspondant à diverses valeurs de n et p. On notera que la distribution est symétrique si $p = 1/2$ et le devient approximativement sinon, dès que n est assez élevé.

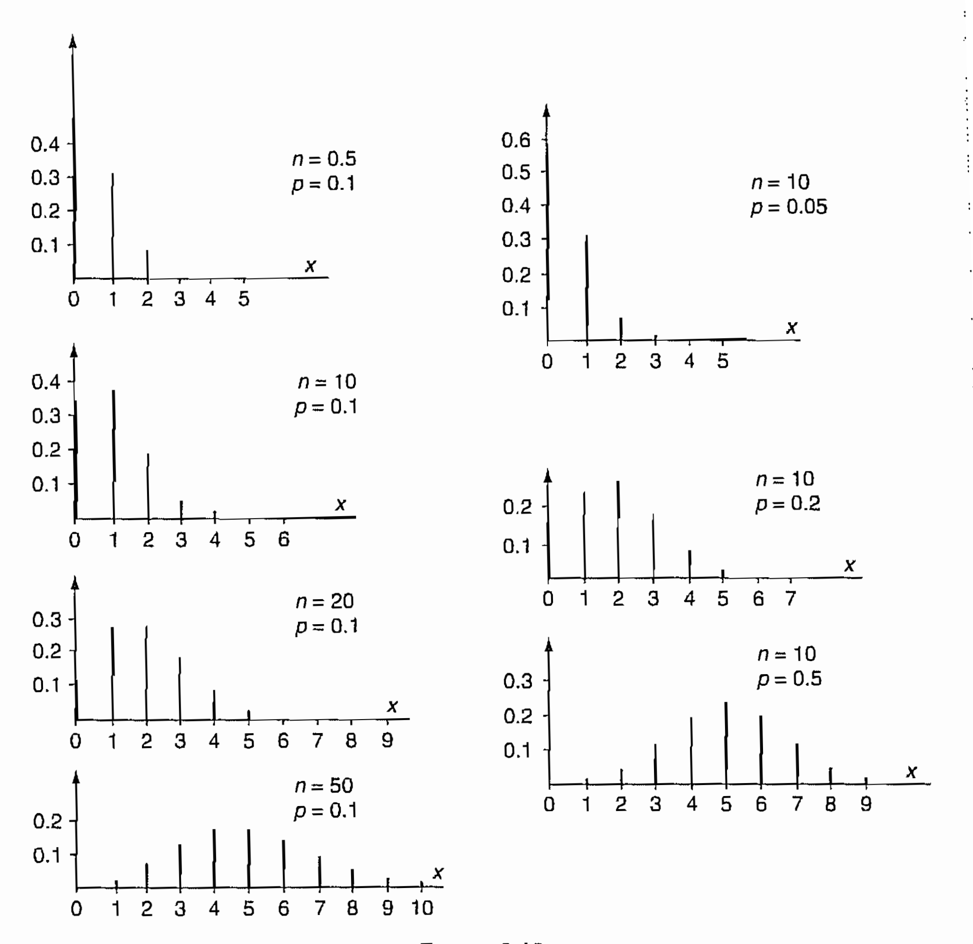


Figure 2.13

Un résultat utile pour l'utilisation des tables : si X suit une loi binomiale $\mathcal{B}(n\,;p)$, $n - X$ suit alors une loi binomiale $\mathcal{B}(n\,;1-p)$.

Pour n grand, on verra plus loin que la loi binomiale peut être approximée soit par une loi de Poisson (si p est petit) soit par une loi de Gauss.

La somme de deux variables aléatoires binomiales indépendantes et de même paramètre p est une variable aléatoire binomiale :

$$\left.\begin{array}{l} X_1 = \mathcal{B}(n_1, p) \\ X_2 = \mathcal{B}(n_2, p) \end{array}\right\} \Rightarrow X_1 + X_2 = \mathcal{B}(n_1 + n_2, p)$$

▬ Démonstration

X_1 : somme de n_1 variables de Bernoulli ;
X_2 : somme de n_2 variables de Bernoulli.

$X_1 + X_2$, somme de $n_1 + n_2$ variables de Bernoulli est bien une variable binomiale d'effectif égal à la somme des effectifs.

Condition nécessaire et suffisante : X_1 et X_2 doivent être indépendantes.

2.2.4 Loi de Poisson $\mathcal{P}(\lambda)$

C'est la loi d'une variable aléatoire entière positive ou nulle qui satisfait à :

$$P(X = x) = \exp(-\lambda)\frac{\lambda^x}{x!} \quad \text{où} \quad x \in \mathbb{N}$$

On peut vérifier tout d'abord qu'il s'agit bien d'une loi de probabilité :

$$\sum_{x=0}^{\infty} P(X = x) = \exp(-\lambda)\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = \exp(-\lambda)\exp(\lambda) = 1$$

À la figure 2.12, quelques diagrammes en bâtons correspondent à diverses valeurs de λ :

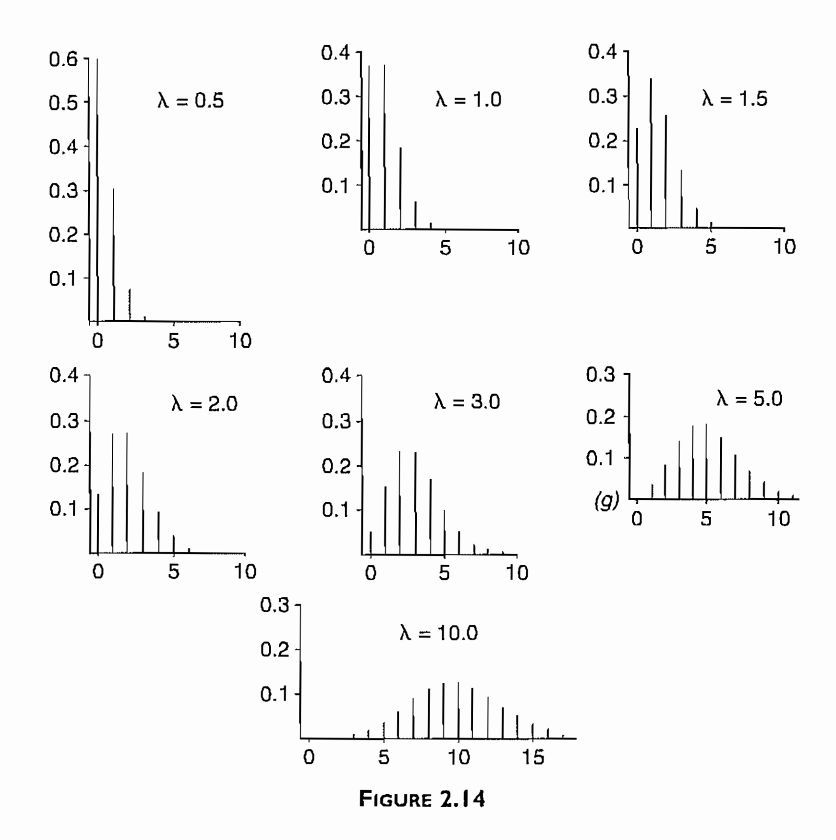


FIGURE 2.14

Le paramètre λ représente à la fois l'espérance et la variance de X.

On obtient la loi de Poisson comme approximation de la loi binomiale dans le schéma suivant :

Soit un événement A de probabilité p très faible (en pratique $p < 0.1$) que l'on essaie d'obtenir quelques fois en répétant l'expérience un grand nombre de fois (en pratique $n > 50$). Le nombre de réalisations de A suit une loi binomiale $\mathscr{B}(n\,;p)$ telle qu'en pratique :

$$\mathscr{B}(n, p) \sim \mathscr{P}(np)$$

c'est-à-dire :

$$C_n^k p^k (1-p)^{n-k} \simeq \exp(-np)\frac{(np)^k}{k!}$$

Nous allons, en fait, établir ce résultat sous la forme mathématique suivante :

THÉORÈME

Soit X_n une suite de variables binomiales $\mathscr{B}(n, p)$ telles que $n \to \infty$ et $p \to 0$ de manière à ce que le produit np tende vers une limite finie λ. Alors la suite de variables aléatoires X_n converge en loi vers une variable de Poisson $\mathscr{P}(\lambda)$.

Les notions de convergence seront étudiées en détail au paragraphe 2.7.

Démonstration

$$C_n^x p^x (1-p)^{n-x} = \frac{n(n-1)\dots(n-x+1)}{x!} p^x (1-p)^{n-x}$$

$$= \frac{(pn)^x}{x!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\dots\left(1-\frac{x-1}{n}\right)(1-p)^{n-x}$$

Faisons tendre $n \to \infty$. Tous les termes $\left(1-\frac{1}{n}\right)\dots\left(1-\frac{x-1}{n}\right)$ tendent vers 1, leur produit tend vers 1 car ils sont en nombre fini.

Décomposons $(1-p)^{n-x}$ en $(1-p)^n(1-p)^{-x}$

$$(1-p)^{-x} \to 1 \text{ car } p \to 0.$$

Quant à $(1-p)^n \sim \left(1-\frac{\lambda}{n}\right)^n$ il tend vers $\exp(-\lambda)$ donc :

$$C_n^x p^x (1-p)^{n-x} \to \left(\frac{np}{x!}\right)^x \exp(-\lambda) \qquad \text{c.q.f.d.}$$

La suite des espérances des binomiales X_n : $E(X_n) = np$ converge vers λ :

$$\boxed{E(X) = \lambda}$$

En effet : $$E(X) = \sum_{x=0}^{\infty} \exp(-\lambda)\frac{\lambda^x}{x!} = \sum_{x=1}^{\infty} \exp(-\lambda)\frac{\lambda^x}{(x-1)!}$$

car le premier terme est nul :

$$E(X) = \exp(-\lambda)\lambda \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \exp(-\lambda)\lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$$

$$= \exp(-\lambda)\lambda \exp(\lambda) = \lambda$$

La suite des variances des binomiales X_n : $V(X_n) = np(1-p)$ tend aussi vers λ car $np \to \lambda$, $p \to 0$.

Montrons que $\boxed{V(X) = \lambda}$.

Démonstration

$$V(X) = E(X^2) - [E(X)]^2 = E(X^2) - \lambda^2$$

$$E(X^2) = \sum_{x=0}^{\infty} x^2 \exp(-\lambda)\frac{\lambda^x}{x!} = \sum_{x=1}^{\infty} x \exp(-\lambda)\frac{\lambda^x}{(x-1)!}$$

avec $x = x - 1 + 1$, il vient :

$$E(X^2) = \sum_{x=2}^{\infty} \exp(-\lambda)\frac{\lambda^x}{(x-2)!} + \sum_{x=1}^{\infty} \exp(-\lambda)\frac{\lambda^x}{(x-1)!}$$

$$E(X^2) = \lambda^2 \exp(-\lambda) \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda \exp(-\lambda) \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$E(X^2) = \lambda^2 \exp(-\lambda)\exp(\lambda) + \lambda \exp(-\lambda)\exp(\lambda) = \lambda^2 + \lambda$$

donc $V(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Donc $\sigma = \sqrt{\lambda}$.

On verra plus loin que la somme de deux variables de Poisson indépendantes est encore une variable de Poisson. Lorsque λ est grand, on verra que la loi de Poisson peut être approximée par la loi de Gauss.

La loi de Poisson s'obtient aussi comme loi exacte du nombre d'événements survenant pendant une période donnée, sous certaines conditions (voir plus loin le paragraphe consacré au processus de Poisson).

Exemples d'application de la loi de Poisson :

- loi du nombre de suicides par an dans un pays donné ;
- loi du nombre d'appels téléphoniques pendant un intervalle de temps T ;
- loi du nombre de pièces défectueuses dans une livraison importante, la production étant de bonne qualité ;
- etc.

2.2.5 Loi hypergéométrique $\mathcal{H}(N, n, p)$ ou du tirage exhaustif

Soit une population de N individus parmi lesquels une proportion p (donc Np individus) possède un certain caractère. On prélève un échantillon de n individus parmi cette population (le tirage pouvant s'effectuer d'un seul coup ou au fur et à mesure mais sans remise). Soit X le nombre aléatoire d'individus de l'échantillon possédant la propriété envisagée. X suit la loi hypergéométrique et l'on a :

$$\boxed{P(X = x) = \frac{C_{Np}^{x} C_{N-Np}^{n-x}}{C_N^n}}$$

min X = max $(0 ; n - Nq)$;

max X = min $(n ; Np)$;

C_N^n nombre d'échantillons possibles ;

C_{Np}^x nombre de groupes de x individus possédant la propriété ;

C_{N-Np}^{n-x} nombre de groupes de $(n - x)$ individus ne possédant pas la propriété.

Le nombre n/N est appelé taux de sondage.

On peut considérer X comme une somme de n variables de Bernoulli $X_1, X_2, \ldots, X_n$ non indépendantes correspondant aux tirages successifs de n individus.

Nous allons montrer que ces variables X_i ont toutes le même paramètre égal à p.

On sait que $E(X_1) = P(X_1 = 1)$ et il est évident que $P(X_1 = 1) = p$.

Cherchons $E(X_2) = P(X_2 = 1)$. Comme X_2 et X_1 sont liés, on a :

$$P(X_2 = 1) = P(X_2 = 1|X_1 = 1)P(X_1 = 1) + P(X_2 = 1|X_1 = 0)P(X_1 = 0)$$

soit :

$$P(X_2 = 1) = \frac{Np - 1}{N - 1}p + \frac{Np}{N - 1}(1 - p)$$

$$= \frac{Np^2 - p + Np - Np^2}{N - 1} = p\frac{(N - 1)}{N - 1} = p$$

De même $E(X_3) = E(X_4) = \cdots = p$.

2.2.5.1 *Espérance de l'hypergéométrique*

$$E(X) = E(\Sigma X_i) = \Sigma E(X_i) \qquad \boxed{E(X) = np}$$

L'espérance ne dépend pas de N et est la même que dans le cas du tirage avec remise (loi binomiale).

2.2.5.2 *Variance de l'hypergéométrique*

Comme il n'y a pas indépendance :

$$V(X) = \sum V(X_i) + 2\sum_{i>j}\text{cov}(X_i, X_j) = \sum V(X_i) + \sum_{i\neq}\text{cov}(X_i, X_j)$$

le terme $\Sigma V(X_i)$ vaut $np(1-p)$ (terme binomial).

On a : $$\text{cov}(X_i, X_j) = E(X_iX_j) - p^2 = P(X_iX_j = 1) - p^2$$

et : $$P(X_iX_j = 1) = P(X_j = 1|X_i = 1)P(X_i = 1) = P(X_j = 1|X_i = 1)p$$

$P(X_j = 1|X_i = 1)$ ne dépend pas des indices i et j et vaut par exemple $P(X_2 = 1|\, X_1 = 1) = \dfrac{Np-1}{N-1}$.

Donc : $$\text{cov}(X_i, X_j) = p\,\frac{Np-1}{N-1} - p^2$$

Comme il y a $n(n-1)$ manières de prendre des couples (X_i et X_j), il vient :

$$V(X) = np(1-p) + n(n-1)\left[p\,\frac{Np-1}{N-1} - p^2\right]$$

soit : $$\boxed{V(X) = \frac{N-n}{N-1}np(1-p)}$$

2.2.5.3 *Tendance vers la loi binomiale*

Si $N \to \infty$, $\mathcal{H}(N, n, p)$ tend vers $\mathcal{B}(n, p)$.

Démonstration

$$\frac{C_{Np}^{x}C_{N-Np}^{n-x}}{C_N^n} = \frac{Np!}{(Np-x)!x!}\,\frac{(N(1-p))!}{(n-x)!(N-Np-n+x)!}\,\frac{n!(N-n)!}{N!}$$
$$= C_n^x\,\frac{Np!}{(Np-x)!}\,\frac{Nq!}{(Nq-n+x)!}\,\frac{(N-n)!}{N!}$$

avec $q = 1-p$.

$$\frac{Np!}{(Np-x)!} = \frac{1\cdot2\cdot3\cdots Np}{1\cdot2\cdot3\cdots(Np-x)} = Np(Np-1)\ldots(Np-x+1)$$

Si N est grand, $Np-1 \sim Np-2 \cdots \sim (Np-x+1) \sim Np$ car x est négligeable devant Np.

Donc : $$\frac{Np!}{(Np-x)!} \sim (Np)^x$$

De même : $$\frac{Nq!}{(Nq-n+x)!} \sim (Nq)^{n-x} \quad \text{et} \quad \frac{N!}{(N-n)!} \sim N^n$$

donc : $$\frac{C_{Np}^{x}C_{Nq}^{n-x}}{C_N^n} \sim C_n^x\,\frac{(Np)^x(Nq)^{n-x}}{N^n} = C_n^x p^x q^{n-x} \qquad \text{c.q.f.d.}$$

En pratique, ce résultat s'applique dès que $n/N < 10\ \%$, c'est-à-dire dès que la population est 10 fois plus grande que l'échantillon, ce qui arrive fréquemment en sondages.

Un échantillon de 2000 individus conviendra donc aussi bien pour faire un sondage dans une ville de 200 000 habitants que dans une ville de 2 millions d'habitants.

2.2.6 Lois géométrique, de Pascal, binomiale négative

La ***loi géométrique*** est la loi du nombre d'essais nécessaires pour faire apparaître un événement de probabilité p :

$$P(X = x) = p(1 - p)^{x-1} \qquad x = 1, 2, \ldots, \infty$$

En posant $q = 1 - p$, on trouve aisément :

$$E(X) = \frac{1}{p} \qquad V(X) = \frac{q}{p^2} \qquad \gamma_1 = \frac{2 - p}{q} \qquad \gamma_2 = 9 + \frac{p^2}{q}$$

La ***loi de Pascal*** d'ordre n est la loi du nombre d'essais nécessaires pour observer n fois un événement A de probabilité p. L'expérience devant se terminer par A, on a :

$$P(X = x) = pC_{x-1}^{n-1}\, p^{n-1} q^{(x-1)-(n-1)} = C_{x-1}^{n-1} p^n q^{x-n} \qquad \text{pour } x = n, n + 1, \ldots, \infty$$

Cette loi est la somme de n lois géométriques indépendantes (apparition de A pour la première fois, puis pour la deuxième fois, etc.), on a :

$$E(X) = \frac{n}{p} \qquad V(X) = \frac{nq}{p^2} \qquad \gamma_1 = \frac{2 - p}{\sqrt{nq}} \qquad \gamma_2 = 3 + \frac{p^2 + 6q}{nq}$$

La ***loi binomiale négative*** est la loi de $Y = X - n$:

$$P(Y = y) = C_{n+y-1}^{n-1}\, p^n q^y$$

Son nom vient du fait suivant : en posant $Q = 1/p$, $P = (1 - p)/p$, on a :

$$P(Y = y) = C_{n+y-1}^{n-1}\, P^y Q^{-n-y}$$

terme général du développement de $(Q - P)^{-n}$ d'où :

$$E(X) = nP \quad V(Y) = nPQ \qquad \gamma_1 = \frac{P + Q}{\sqrt{nPQ}} \qquad \gamma_2 = 3 + \frac{1 + 6PQ}{nPQ}$$

que l'on comparera aux moments de la binomiale $\mathcal{B}(n, p)$.

2.3 DISTRIBUTIONS CONTINUES USUELLES

2.3.1 Loi uniforme sur [0, *a*]

Sa densité est :

$f(x) = \dfrac{1}{a}$ sur $[0, a]$;

$f(x) = 0$ ailleurs ;

$F(x) = \frac{x}{a}$ sur $[0, a]$;

$F(x) = 0$ sur $[-\infty, 0]$; $F(x) = 1$ sur $[a, +\infty]$ (voir fig. 2.13).

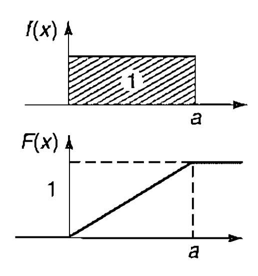


FIGURE 2.15

Son espérance vaut $E(X) = \frac{a}{2}$ car la densité est symétrique.

Sa variance vaut $V(X) = \int_0^a x^2 \frac{1}{a} \mathrm{d}x - \frac{a^2}{4} = \frac{a^2}{12}$.

La somme de deux lois uniformes n'est pas une loi uniforme. Ainsi, soit X et Y deux variables uniformes sur $[0, a]$; leur somme Z, si elles sont indépendantes, est une variable de densité triangulaire (fig. 2.16).

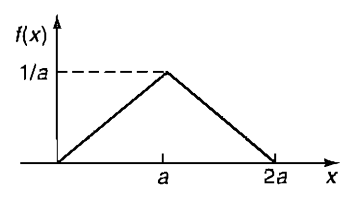


FIGURE 2.16

2.3.2 Loi exponentielle de paramètre λ

Sa densité est $f(x) = \lambda \exp(-\lambda x)$ si $x > 0$.

On trouve sans difficulté :

$\boxed{E(X) = 1/\lambda}$ $\boxed{V(X) = 1/\lambda^2}$ $\boxed{\gamma_1 = 2}$ $\boxed{\gamma_2 = 9}$

En fiabilité, cette loi est très utilisée pour représenter la durée de vie de circuits électroniques. L'espérance $1/\lambda$ est souvent appelée le MTBF (*Mean Time Between Failure*) et λ le taux de défaillance car $h(x) = \frac{f(x)}{1 - F(x)} = \lambda$ et est constant.

2.3.3 Lois gamma

La loi exponentielle est un cas particulier d'une famille de lois appelés lois γ. Précisément, si X est une loi exponentielle de paramètre λ, λX est une variable suivant une loi γ_1.

On dit qu'une variable aléatoire positive X suit une loi gamma de paramètre r, notée γ_r si sa densité est donnée par :

$$\boxed{f(x) = \frac{1}{\Gamma(r)} \exp(-x)\, x^{r-1}}$$

Il s'agit bien d'une densité car $f(x)$ est > 0 et $\int_0^\infty f(x)\,\mathrm{d}x = 1$ par définition de $\Gamma(r)$. Les lois γ_r avec r entier > 1 sont aussi connues sous le nom de lois d'Erlang.

2.3.3.1 Espérance

$$\boxed{E(X) = r}$$

En effet :

$$E(X) = \frac{1}{\Gamma(r)} \int_0^\infty x^r \exp(-x)\,\mathrm{d}x = \frac{\Gamma(r+1)}{\Gamma(r)} = r$$

2.3.3.2 Variance

$$\boxed{V(X) = r}$$

En effet :

$$V(X) = E(X^2) - [E(X)]^2 = \frac{1}{\Gamma(r)} \int_x^\infty x^{r+1} \exp(-x)\,\mathrm{d}x - r^2$$

soit :

$$V(X) = \frac{\Gamma(r+2)}{\Gamma(r)} - r^2 = (r+1)\frac{\Gamma(r+1)}{\Gamma(r)} - r^2 = r(r+1) - r^2$$

Cette loi présente donc une certaine analogie avec la loi de Poisson mais en continu. Les courbes de densité sont représentées à la figure 2.17.

Les lois γ vérifient la propriété d'additivité suivante :

THÉORÈME

Si X et Y sont des variables indépendantes suivant respectivement des lois γ_r et γ_s, alors $X + Y$ suit une loi γ_{r+s}.

Ce résultat sera démontré au paragraphe 2.5 de ce chapitre.

Les lois γ sont liées aux lois du χ^2 utilisées en statistique par une formule simple (voir chapitre 4) :

Si X suit une loi γ_r, $2X$ suit une loi χ^2_{2r}.

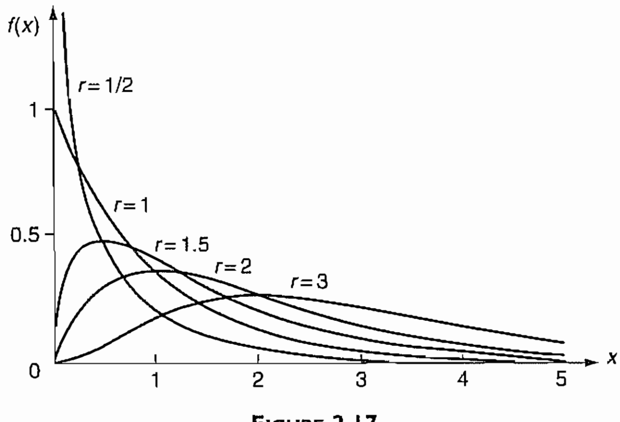


FIGURE 2.17

2.3.4 Lois bêta

2.3.4.1 *Loi bêta de type I*

C'est la loi d'une variable X ; $0 \leq X \leq 1$ dépendant de deux paramètres n et p dont la densité est :

$$f(x) = \frac{1}{B(n,p)} x^{n-1}(1-x)^{p-1} \qquad n, p > 0 \qquad \text{où } B(n,p) = \frac{\Gamma(n)\,\Gamma(p)}{\Gamma(n+p)}$$

On trouve :

$$\boxed{E(X) = \frac{n}{n+p}} \qquad \boxed{V(X) = \frac{np}{(n+p+1)(n+p)^2}}$$

Ces lois sont utilisées en statistique bayésienne pour représenter la distribution *a priori* de la probabilité d'un événement.

L'allure de quelques courbes de densité est donnée par la figure 2.18.

2.3.4.2 *Loi bêta de type II*

Soit X une variable suivant une loi bêta $\mathrm{I}(n, p)$; alors, par définition, $Y = X/(1-X)$ suit une loi bêta de type II dont la densité s'obtient aisément par changement de variable :

$$f(y) = \frac{1}{B(n,p)} \frac{y^{n-1}}{(1+y)^{n+p}} \qquad 0 \leq Y \leq \infty$$

$$\boxed{E(Y) = \frac{n}{p-1}} \qquad \boxed{V(Y) = \frac{n(n+p-1)}{(p-1)^2(p-2)}}$$

PROPRIÉTÉ

Le rapport de deux variables indépendantes suivant des lois γ_n et γ_p respectivement suit une loi bêta $\mathrm{II}(n, p)$.

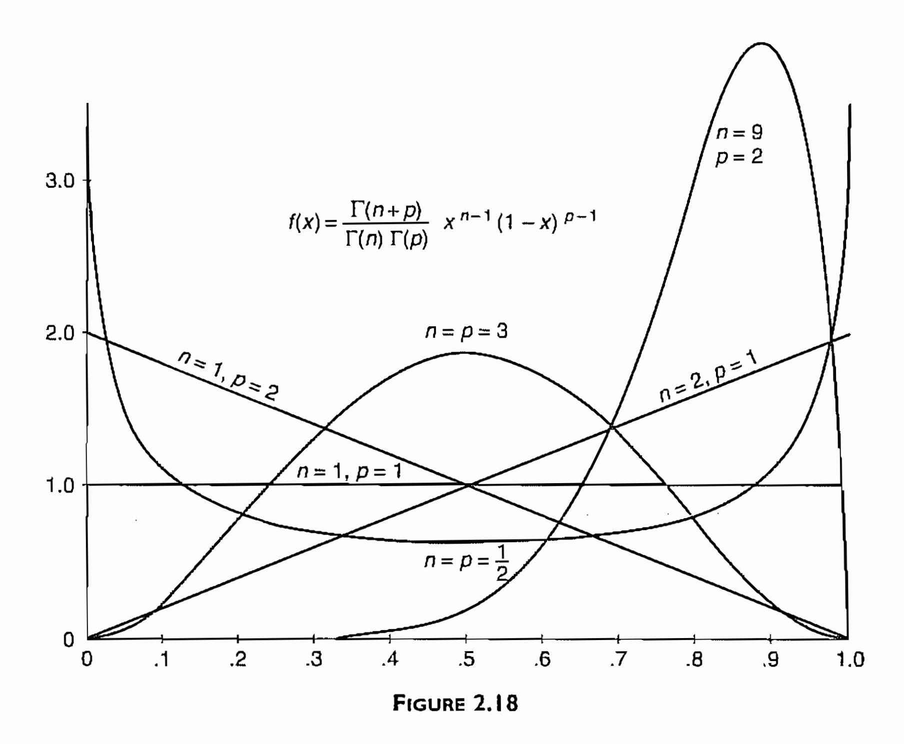


FIGURE 2.18

La démonstration est laissée au soin du lecteur.

Les diverses valeurs de n et p font que cette loi s'adapte bien à la représentation de nombreux phénomènes aléatoires positifs (temps d'attente, durées de vie, méthode Pert avec durée aléatoire).

Ces lois sont liées aux lois de Fisher-Snedecor utilisées en statistique (voir chapitre 4).

2.3.4.3 Loi de l'arc sinus

La loi bêta I (1/2 ; 1/2) dont la densité est $f(x) = \dfrac{1}{\pi\sqrt{x(1-x)}}$ porte le nom de loi de l'arc sinus car sa fonction de répartition est :

$$F(x) = \frac{2}{\pi} \text{arc sin}(\sqrt{x})$$

On a $E(X) = 1/2$, $V(X) = 1/8$, $\gamma_1 = 0$, $\gamma_2 = 1.5$.

Cette loi assez paradoxale, puisque l'espérance est la valeur la moins probable et les valeurs extrêmes sont les plus probables, s'applique en particulier dans certains phénomènes liés aux jeux de hasard.

Par exemple, deux joueurs jouent à un jeu équitable (du type pile ou face). Soit $S_1, S_2, \ldots S_n$ la suite des gains d'un des deux joueurs ; si X désigne la proportion du temps passé en gain positif, la loi limite de X quand $n \to \infty$ est la loi de l'arc sinus. Il y a donc plus de chance d'être constamment en gain ou constamment en perte que d'être dans le cas médian (c'est la loi de la persistance de la chance ou de la malchance . . .).

Cette loi a pu être appliqué à la persistance du temps en météorologie et rend compte du fait qu'il est plus fréquent de battre des records (de froid ou de chaud) que d'avoir un temps moyen.

2.3.5 La loi de Laplace-Gauss

Cette loi joue un rôle fondamental en probabilités et statistique mathématique. Elle constitue un modèle fréquemment utilisé dans divers domaines : variation du diamètre d'une pièce dans une fabrication industrielle, répartition des erreurs de mesure autour de la « vraie valeur », etc.

Malgré son appellation malencontreuse de loi normale[1], elle est cependant loin de décrire tous les phénomènes physiques et il faut se garder de considérer comme anormale une variable ne suivant pas la loi de Laplace-Gauss. Son rôle principal en statistique provient en réalité de ce qu'elle apparaît comme loi limite de caractéristiques liées à un échantillon de grande taille. Le théorème central-limite que nous établirons au paragraphe 2.7 montre que dans certaines conditions la somme, et donc la moyenne, de variables indépendantes et de même loi est asymptotiquement une loi normale.

X suit une loi normale LG(m ; σ) si sa densité est[2] :

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)}$$

Par suite de la symétrie de f et comme l'intégrale de X converge, $\boxed{E(X) = m}$.

Avec le changement de variable aléatoire $\boxed{U = \frac{X-m}{\sigma}}$, on trouve que la densité de U est :

$$f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$$

U est une LG (0, 1), donc toute variable X LG (m ; σ) se ramène simplement à la variable U par $X = m + \sigma U$.

Montrons que $V(U) = 1$:

$$V(U) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} u^2 \exp\left(-\frac{1}{2}u^2\right) du = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} u^2 \exp\left(-\frac{1}{2}u^2\right) du$$

Posons $t = u^2/2$, il vient $u\, du = dt$:

$$V(U) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} \exp(-t)\, dt = \frac{2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) = \frac{2}{\sqrt{\pi}} \frac{1}{2} \Gamma\left(\frac{1}{2}\right)$$

1 Cette dénomination fut introduite par K. Pearson qui voulait éviter les querelles d'antériorité concernant son introduction en statistique et l'a d'ailleurs regretté par la suite comme l'indique cette citation : *Many years ago I called the Laplace-Gaussian curve the normal curve which name, while it avoids an international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another 'abnormal.' That belief is, of course, not justifiable. It has led many writers to try and force all frequency by aid of one or another process of distortion into a 'normal' curve* (paper read to the Society of Biometricians and Mathematical Statisticians, June 14, 1920).

2 La notation LG sera utilisée couramment dans cet ouvrage. La notation N(m ; σ) sera également utilisée.

comme $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$: $\boxed{V(U) = 1}$.

Il en résulte que σ est l'écart-type de X.

La fonction de répartition et la densité de X sont représentées sur la figure 2.19.

Les points d'inflexion sont à $\pm\sigma$ de part et d'autre de m.

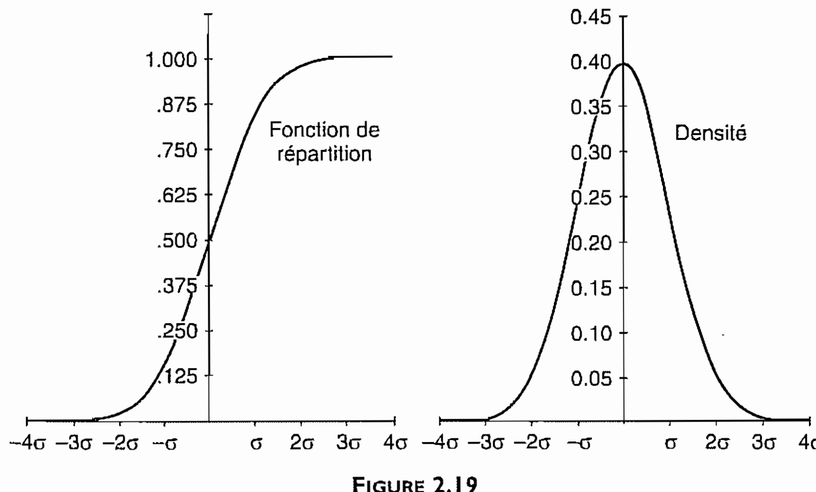


FIGURE 2.19

2.3.5.1 *Valeurs remarquables*

$$\boxed{\begin{aligned} P(m - 1.64\sigma < X < m + 1.64\sigma) &= 0.90 \\ P(m - 1.96\sigma < X < m + 1.96\sigma) &= 0.95 \\ P(m - 3.09\sigma < X < m + 3.09\sigma) &= 0.998 \end{aligned}}$$

2.3.5.2 *Moments*

Ils existent pour tout ordre.

Par suite de la symétrie, tous les moments d'ordre impair sont nuls. Calculons les moments d'ordre pair :

$$\mu_{2k} = \int_{\mathbb{R}} u^{2k} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} u^{2k} \exp\left(-\frac{1}{2}u^2\right) du$$

Posons $y = u^2/2$:

$$\mu_{2k} = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} (2y)^k \exp(-y) \frac{dy}{\sqrt{2y}} = \frac{2^k}{\sqrt{\pi}} \int_0^{\infty} y^{k-\frac{1}{2}} \exp(-y) dy$$

d'où :
$$\mu_{2k} = \frac{2^k}{\sqrt{\pi}}\Gamma\left(k+\frac{1}{2}\right)$$

Comme : $\Gamma\left(k+\frac{1}{2}\right) = \frac{1\cdot 3\cdot 5\cdots(2k-1)}{2^k}\Gamma\left(\frac{1}{2}\right) = \frac{1\cdot 3\cdot 5\cdots 2k-1}{2^k}\sqrt{\pi}$

(voir annexes) il vient :
$$\mu_{2k} = 1\cdot 3\cdots(2k-1) = \frac{(2k)!}{2^k k!}$$

on en déduit $\mu_4 = 3$, d'où $\gamma_2 = 3$.

2.3.5.3 Additivité

Les variables de Gauss possèdent la propriété d'additivité.

THÉORÈME

Si X_1 et X_2 sont des variables indépendantes suivant respectivement des lois LG$(m_1\,;\sigma_1)$ *et* LG$(m_2\,;\sigma_2)$ *alors X_1+X_2 est une variable* LG$(m_1+m_2\,;\sqrt{\sigma_1^2+\sigma_2^2})$.

Ce résultat fondamental sera démontré au paragraphe 2.6 à l'aide des fonctions caractéristiques.

On ne peut cependant pas dire que toute combinaison linéaire de p variables gaussiennes non indépendantes soit encore gaussienne. Il faut pour cela que le p-uple de variables suive une loi normale à p-dimensions (dont c'est précisément la définition. *cf.* chapitre 4).

2.3.5.4 Loi de U^2

D'après la formule établie à la fin du paragraphe 2.1.2.2, la densité de $T = U^2$ est :
$$g(t) = \frac{f(\sqrt{t})}{\sqrt{t}} = \frac{1}{\sqrt{2\pi}}t^{-1/2}\exp\left(-\frac{t}{2}\right)$$

en remplaçant $f(t)$ par $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$, on remarque que $U^2/2$ suit une loi $\gamma_{1/2}$ ou loi du khi-deux à un degré de liberté (voir chapitre 4).

2.3.6 La loi log-normale

C'est la loi d'une variable positive X telle que son logarithme népérien suive une loi de Laplace-Gauss :
$$\ln X \sim \text{LG}(m\,;\sigma)$$

Sa densité s'obtient par un simple changement de variable et on trouve :
$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln x - m}{\sigma}\right)^2\right)$$

$$E(X) = \exp\left(m + \frac{\sigma^2}{2}\right) \qquad V(X) = (\exp(2m + \sigma^2))(\exp \sigma^2 - 1)$$

On utilise parfois la loi log-normale à trois paramètres γ, m, σ telle que :

$$\ln (X - \gamma) \sim LG(m ; \sigma) \qquad \text{avec } X > \gamma.$$

La figure 2.20 représente la densité de la loi log-normale d'espérance 2 et d'écart-type 1 :

$$(m \simeq 0.58 \qquad \sigma \simeq 0.47)$$

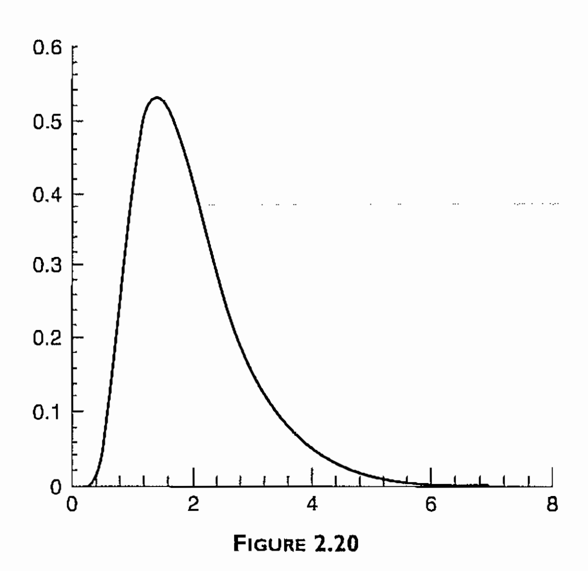


FIGURE 2.20

2.3.7 Loi de Cauchy

C'est la loi d'une variable X réelle de densité :

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

Sa fonction de répartition est $F(x) = \frac{1}{\pi} \operatorname{arc\,tg} x + \frac{1}{2}$.

X ne possède aucun moment fini car l'intégrale $\int_{\mathbb{R}} \frac{x}{\pi(1 + x^2)} dx$ diverge.

On montre que la loi de Cauchy est la loi du rapport de deux variables LG(0 ; 1) indépendantes. Elle s'identifie à T_1 variable de Student de degré 1 (voir chapitre 4).

2.3.8 Loi de Weibull

Très utilisée en fiabilité, la loi de Weibull à deux paramètres donne la probabilité qu'une durée X de fonctionnement sans défaillance soit supérieure à x par :

$$P(X > x) = e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

En d'autres termes, $\left(\frac{X}{\beta}\right)^{\alpha}$ suit une loi exponentielle.

La densité de X est : $f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$

Le paramètre α, qui est sans dimension, est appelé paramètre de forme. Selon ses valeurs, la densité de probabilité est plus ou moins dissymétrique. Le paramètre de forme est lié au vieillissement : quand il vaut 1, on a une loi exponentielle caractéristique des matériels sans usure ni fatigue. Quand il est plus grand que 1, on est en présence de fatigue : le taux instantané de défaillance $h(x)$ est alors croissant avec x :

$$h(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}$$

Si α est inférieur à 1, on a affaire à un matériel qui se bonifie avec le temps.

Le paramètre β s'exprime dans la même unité que X (jours, heures, nombre de cycles, etc.). C'est un paramètre d'échelle lié à la durée de vie médiane par :

$$\beta = \frac{médiane}{(\ln(2))^{\frac{1}{\alpha}}}.$$

La figure 2.21 donne la densité d'une loi de Weibull avec $\alpha = 2$ et $\beta = 1$.

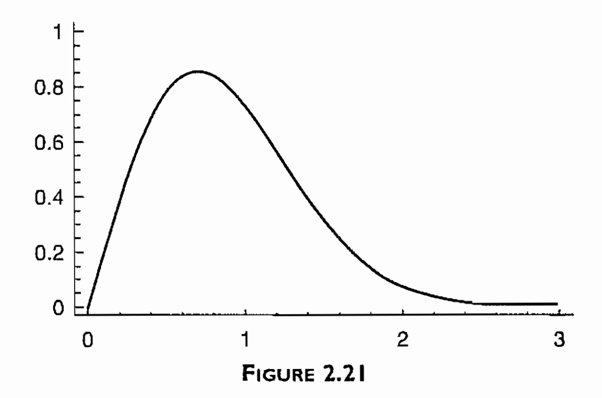


FIGURE 2.21

La relation $E\left[\left(\frac{X}{\beta}\right)^{r}\right] = \Gamma\left(1 + \frac{r}{\alpha}\right)$ permet de calculer les moments de X. Dans l'exemple précédent $\alpha = 2$ et $\beta = 1$, on trouve $E(X) = \frac{\sqrt{\pi}}{2}$ et $V(X) = \frac{3\pi}{4}$ (voir annexe 4).

2.3.9 Loi de Gumbel

Cette loi est utilisée pour les distributions de valeurs extrêmes (voir chapitre 12). Sous sa forme standard sa fonction de répartition est :

$$\boxed{F(x)=\exp(-\exp(-x))}$$

soit : $$\boxed{f(x) = \exp(-x - \exp(-x))} \qquad \text{(fig. 2.22)}$$

$\exp(-X)$ suit donc une loi γ_1.

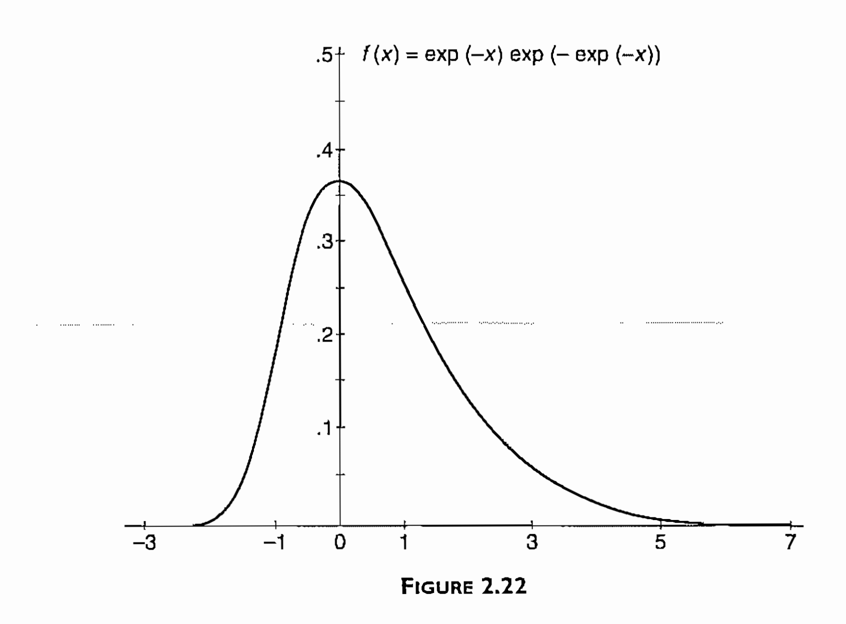


FIGURE 2.22

Ses moments sont :

$$E(X) = 0.57722\ldots \qquad \text{(constante d'Euler)}$$

$$V(X) = \frac{\pi^2}{6}$$

$$\gamma_1 = 1.29857$$

$$\gamma_2 = 5.4$$

La loi de Gumbel est utilisée pour modéliser des phénomènes tels que : crue maximale annuelle d'une rivière, magnitude du plus grand tremblement de terre enregistré en une année, etc.

2.4 LE PROCESSUS PONCTUEL DE POISSON

Considérons une famille X_t de variables de Bernoulli ($X_t = 1$ si un événement (arrivée d'un client, accident, appel téléphonique . . .) se produit à l'instant t) : on s'intéressera à la répartition des dates d'arrivée des événements, ainsi quà N_t nombre d'événements entre 0 et t.

2.4.1 Flux poissonnien d'événements

Un processus de Poisson représente l'apparition d'événements aléatoires E_1, E_2 . . ., E_n, etc., satisfaisant aux trois conditions suivantes :

- Les temps d'attente entre deux événements E_1 E_2, E_2 E_3, etc. sont des variables indépendantes (processus sans mémoire).
- La loi du nombre d'événements arrivant dans l'intervalle [t ; t + T] ne dépend que de T. *Si* $T = 1$, on notera c son espérance, dite « cadence ».
- Deux événements ne peuvent arriver simultanément.

Soit $p_0(h)$ la probabilité qu'aucun événement ne se produise pendant une durée h ; d'après la deuxième condition, $p_0(h)$ ne dépend que de h et non de l'instant considéré.

Soient trois instants t, $t + h$, $t + h + k$. La probabilité qu'il ne se passe rien entre t et $t + h + k$ est $p_0(h + k)$; d'après l'axiome d'indépendance, on a :

$$p_0(h + k) = p_0(h)\,p_0(k) \qquad \forall h, \forall k$$

D'où le résultat :

$$p_0(h) = \exp(-ch) \qquad \text{avec } c > 0$$

Nous montrerons par la suite que c est bien la cadence du phénomène.

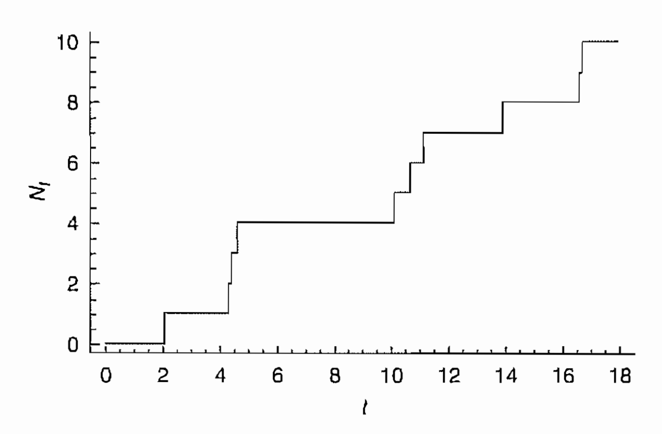


FIGURE 2.23 Une trajectoire d'un processus de Poisson avec $c = 1$; en ordonnée le nombre cumulé d'événements depuis $t = 0$.

2.4.2 Étude de la durée T séparant deux événements consécutifs E_i et E_{i+1}

Soit T cette durée qui est une variable aléatoire, la probabilité que $T > t$ est égale à la probabilité qu'il n'arrive rien pendant une durée t soit :

$$P\,(T > t) = \exp(-ct)$$

d'où la fonction de répartition de T : $P(T < t) = 1 - \exp(-ct)$. La densité vaut alors $f(t) = \exp(-ct)c$ il s'ensuit que cT suit une loi γ_1, donc $E(T) = 1/c$.

2.4.3 Étude de la durée Y séparant $n + 1$ événements

Y est une variable aléatoire somme de n variables indépendantes de même loi :

$$Y = T_1 + T_2 + \cdots + T_n$$

soit :

$$cY = cT_1 + cT_2 + \cdots + cT_n \qquad \text{(fig. 2.24)}$$

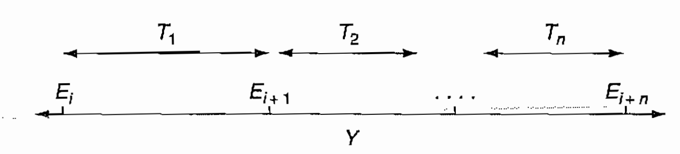


FIGURE 2.24

donc cY suit une loi γ_n ; la densité de Y est :

$$\boxed{f_n(y) = \exp(-cy)\frac{(cy)^{n-1}}{(n-1)!}c}$$

2.4.4 Étude du nombre d'événements se produisant pendant une période de durée T fixée

THÉORÈME

Le nombre d'événements suit une loi de Poisson de paramètre cT.

Démonstration : Soit AB la période d'étude (fig. 2.25) :

A E1 E2 En B

FIGURE 2.25

On a la relation évidente : $P(N = n) = P(N \geq n) - P(N \geq n + 1)$.

La probabilité $P(N \geq n)$ est aussi la probabilité que la durée AE_n soit inférieure à T ; cette durée est constituée de $AE_1 + E_1E_2 + \cdots E_{n-1}E_n$ qui sont des lois exponentielles indépendantes ; donc $c\,AE_n$ suit une loi γ_n et l'on a :

$$P(N = n) = \int_0^T \exp(-ct)\frac{(ct)^{n-1}}{(n-1)!}c\,dt - \int_0^T \exp(-ct)\frac{(ct)^n}{n!}c\,dt$$

En intégrant par parties la première intégrale, il vient :

$$\int_0^T \exp(-ct)\,\frac{(ct)^{n-1}}{(n-1)!}\,c\,\mathrm{d}t = \int_0^T \exp(-ct)\,\mathrm{d}\left(\frac{(ct)^n}{n!}\right)$$

$$= \exp(-cT)\,\frac{(cT)^n}{n!} + \int_0^T \exp(-ct)\,\frac{(ct)^n}{n!}\,c\,\mathrm{d}t$$

donc :

$$\boxed{P(N = n) = \exp(-cT)\frac{(cT)^n}{n!}}$$

$E(N) = cT$, en particulier si $T = 1$.

On trouve $E(N) = c$; c est donc bien la cadence définie au début de cette partie.

Application importante : Relation entre loi de Poisson et loi du χ^2

Si N suit une loi $\mathcal{P}(\lambda)$ on a :

$$\boxed{P(N \leq n) = P(\chi^2_{2(n+1)} > 2\lambda)}$$

Il suffit de considérer un processus de Poisson de cadence $c = 1$, observé sur une durée λ :

$$P(N \leq n) = P(T_1 + T_2 + \cdots + T_{n-1} > \lambda) = P(\gamma_{n+1} > \lambda)$$

$$= P(2\gamma_{n+1} > 2\lambda) = P(\chi^2_{2(n+1)} > 2\lambda)$$

2.4.5 Étude de la répartition des dates $E_1, E_2, \ldots E_n$ dans l'intervalle AB

Posons $A = 0$ et cherchons la loi de probabilité conjointe des dates $E_1, E_2, \ldots, E_n$ et de N nombre d'événements survenus.

La probabilité pour que le premier événement se passe entre t_1 et $t_1 + \mathrm{d}t_1$ est : $c\exp(-ct_1)\,\mathrm{d}t_1$.

La probabilité conditionnelle que E_2 arrive entre t_2 et $t_2 + \mathrm{d}t_2$ sachant E_1 est : $c\exp(-c(t_2 - t_1))\,\mathrm{d}t_2$, etc.

La probabilité qu'aucun événement n'arrive après E_n sachant la date de E_n est : $\exp(-c(T - t_n))$; d'où :

$$f(t_1, t_2, \ldots, t_n, n) = c^n \exp(-cT)$$

La loi conditionnelle :

$$f(t_1, t_2, \ldots, t_n / N = n) = \frac{c^n \exp(-cT)}{\exp(-cT)\,\dfrac{(cT)^n}{n!}} = \frac{n!}{T^n}$$

ce qui prouve que les instants $t_1, t_2, \ldots, t_n$ constituent un échantillon ordonné de la loi uniforme sur $[0, T]$: en effet, si l'on s'intéresse seulement aux dates et non à leur ordre, il faut diviser par $n!$ qui est le nombre d'ordres possibles.

2.4.6 Le processus (N_t)

D'après ce qui précède, N_t suit pour tout t une loi de Poisson $\mathcal{P}(ct)$. Comme $E(N_t) = ct = V(N_t)$, ce processus n'est pas stationnaire mais il est à accroissements stationnaires et indépendants puisque $\forall h$, $N_{t+h} - N_t = \mathcal{P}(h)$.

La fonction de covariance de ce processus est facile à obtenir :

- si $s > t$: $C(t, s) = \text{cov}(N_t \,;\, N_s) = \text{cov}(N_t \,;\, N_t + X) = V(N_t) + \text{cov}(N_t \,;\, X)$: or X est une variable indépendante de N_t (accroissements indépendants) donc :
- si $s \geq t$: $C(t\,;\,s) = V(N_t) = ct$; et on trouve de même si $t > s$: $C(t, s) = cs$; d'où :

$$\boxed{C(t\,;\,s) = c \inf(t\,;\,s).}$$

Cette fonction est continue en $t = s$ donc le processus est continu en moyenne quadratique. Cependant, aucune trajectoire n'est continue puisque (N_t) est une fonction aléatoire en escalier (incréments de 1 à chaque événement).

2.5 CONVOLUTION

Un problème courant consiste à trouver la loi de probabilité d'une somme de deux variables indépendantes $Z = X + Y$.

2.5.1 Cas discret

Le théorème des probabilités totales donne la solution du problème :

$$P(Z = z) = \sum_x P(X = x \cap Y = z - x) = \sum_y P(X = z - y \cap Y = y)$$

Lorsque X et Y sont indépendantes, on a :

$$\boxed{P(Z = z) = \sum_x P(X = x)P(Y = z - x)}$$

Sinon, on peut toujours écrire :

$$P(Z = z) = \sum_x P(X = x)P(Y = z - x/X = x)$$

Remarquons que, pour la sommation, x ne prend pas nécessairement toutes les valeurs possibles de X mais uniquement celles compatibles avec l'événement $Z = z$.

Exemple : Soit X et Y, deux variables de Poisson indépendantes de paramètres λ et μ respectivement :

$$P(X = x) = \exp(-\lambda)\frac{\lambda^x}{x!} \qquad P(Y = y) = \exp(-\mu)\frac{\mu^y}{y!}$$

On a donc :

$$P(Z = z) = \sum_{x=0}^{x=z} \exp(-\lambda)\frac{\lambda^x}{x!}\exp(-\mu)\frac{\mu^{z-x}}{(z-x)!}$$

soit en multipliant et divisant par $z!$:

$$P(Z = z) = \frac{\exp(-(\lambda + \mu))}{z!} \sum_{x=0}^{x=z} C_z^x \lambda^x \mu^{z-x}$$

$$= \frac{\exp(-(\lambda + \mu))}{z!} (\lambda + \mu)^z$$

$Z = X + Y$ est donc une variable de Poisson $\mathcal{P}(\lambda + \mu)$. ■

2.5.2 Cas général

La loi de probabilité de $Z = X + Y$ s'obtient grâce au théorème de la mesure image : en effet, la loi de Z n'est autre que la mesure image de P_{XY} par l'application de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $(x, y) \rightarrow x + y$.

Lorsque X et Y sont indépendants, on a donc le résultat suivant :

THÉORÈME

La loi de probabilité de la somme Z de deux variables indépendantes est la mesure image de $P_X \otimes P_Y$ par l'application $(x, y) \rightarrow x + y$ de $\mathbb{R}^2$ dans $\mathbb{R}$.

Notée $P_X * P_Y = P_Z$ (produit de convolution de deux mesures), elle est telle que pour tout borélien B :

$$P_Z(B) = \int_{\mathbb{R}^2} \mathbb{1}_B(x + y)\, dP_X(x) \otimes dP_Y(y)$$

On remarquera le caractère symétrique en x et y de la formule précédente.

En particulier, si X et Y admettent des densités, on a :

$$P_Z(B) = \int_{\mathbb{R}^2} \mathbb{1}_B(x + y) f(x) g(y)\, dx\, dy$$

Posons $x + y = z$, $x = u$ et appliquons le théorème de Fubini :

$$P_Z(B) = \int_{\mathbb{R}^2} \mathbb{1}_B(z) f(u) g(z - u)\, du\, dz$$

$$= \int_{\mathbb{R}} \mathbb{1}_B(z)\, dz \int_{D_X} f(u) g(z - u)\, du$$

D'après la définition des variables continues, on en déduit que Z admet pour densité :

$$\boxed{k(z) = \int_{D_X} f(u)\, g(z - u)\, du = \int_{D_Y} g(y) f(z - y)\, dy}$$

les domaines D_X et D_Y étant les ensembles de valeurs de X et de Y respectivement compatibles avec l'événement $Z = z$.

Par intégration, on en déduit :

$$P(Z < z) = K(z) = \int_{D_X} f(x)\, G(z - x)\mathrm{d}x = \int_{D_Y} g(y)\, F(z - y)\mathrm{d}y$$

Géométriquement, $K(z)$ représente la mesure du domaine hachuré (fig. 2.26).

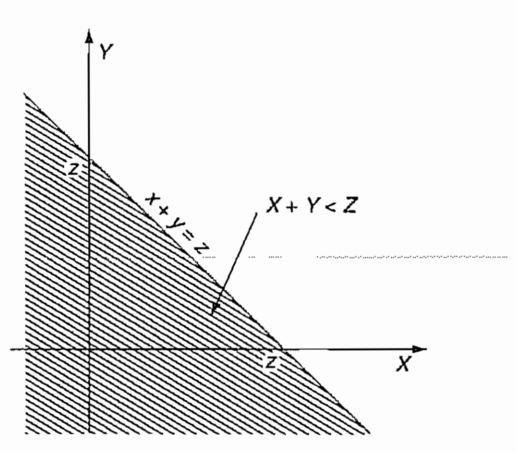


Figure 2.26

2.5.3 Applications

2.5.3.1 Somme de lois γ

Soit X de loi γ_r $f(x) = \dfrac{1}{\Gamma(r)} \exp(-x)\, x^{r-1}$ et Y de loi γ_s $g(y) = \dfrac{1}{\Gamma(s)} \exp(-y)\, y^{s-1}$ indépendante.

$$k(z) = \int_0^z \frac{1}{\Gamma(r)} \exp(-x)\, x^{r-1} \frac{1}{\Gamma(s)} \exp(-(z - x))\, (z - x)^{s-1}\mathrm{d}x$$

$$= \frac{\exp(-z)}{\Gamma(r)\Gamma(s)} \int_0^z x^{r-1}\, (z - x)^{s-1}\mathrm{d}x$$

Posons $x = tz$, il vient :

$$k(z) = \frac{\exp(-z)}{\Gamma(r)\Gamma(s)} \int_0^1 t^{r-1} z^{r-1}\, (z - tz)^{s-1}\mathrm{d}t$$

d'où :

$$k(z) = \frac{\exp(-z) z^{r+s-1}}{\Gamma(r)\Gamma(s)} \int_0^1 t^{r-1}\, (1 - t)^{s-1}\mathrm{d}t$$

$$k(z) = \exp(-z)\, z^{r+s-1} c$$

$k(z)$ étant une densité, la constante c vaut nécessairement $\dfrac{1}{\Gamma(r+s)}$ puisqu'on reconnaît l'expression de la densité d'une loi γ. On en déduit une preuve (probabiliste) de la formule :

$$\int_0^1 t^{r-1}(1-t)^{s-1}\,dt = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$$

Donc si X est une γ_r et Y une γ_s indépendante, $X + Y$ est une γ_{r+s}.

2.5.3.2 Somme de lois uniformes sur [0, 1]

Soient X et Y deux variables continues uniformes sur [0, 1]. La loi de leur somme s'obtient par l'argument géométrique suivant : le couple (X, Y) est uniformément réparti sur le carré unité et l'événement $Z < z$ correspond à la zone hachurée dont il suffit alors de trouver la surface. K et k ont deux déterminations mais sont continues (fig. 2.27).

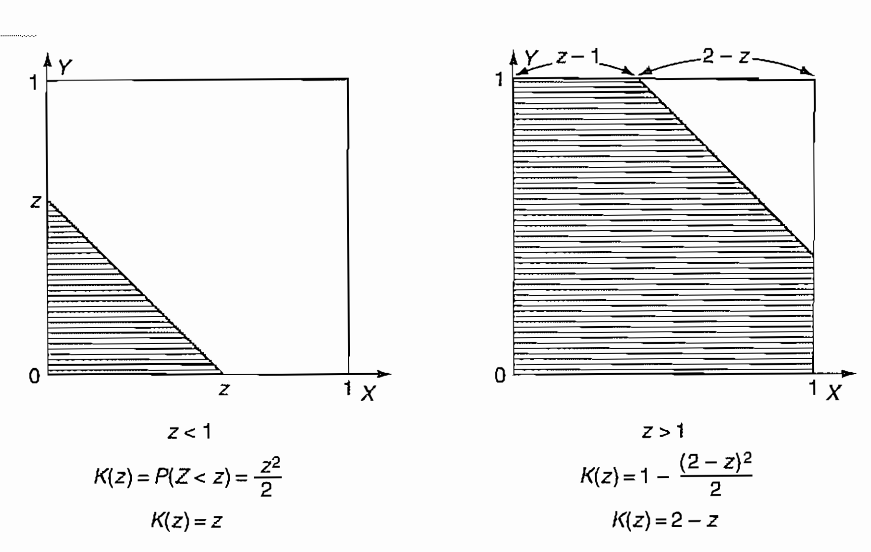


FIGURE 2.27

2.6 FONCTIONS CARACTÉRISTIQUES

2.6.1 Définitions et principales propriétés

2.6.1.1 Définition

La fonction caractéristique d'une variable aléatoire réelle X est la transformée de Fourier de sa loi de probabilité. Elle est notée φ_X et on a :

$$\varphi_X(t) = E[\exp(itX)] = \int_{\mathbb{R}} \exp(itx)\,dP_X(x)$$

Cette fonction ***existe toujours*** car P_X est une mesure bornée et $|\exp(itX)| = 1$. Il s'ensuit que la ***fonction caractéristique est continue***.

Lorsque X possède une densité :

$$\boxed{\varphi_X(t) = \int_{\mathbb{R}} \exp(itx) f(x)\, dx}$$

2.6.1.2 Fonction caractéristique d'une forme linéaire

$$\boxed{\begin{aligned} \varphi_{\lambda X}(t) &= \varphi_X(\lambda t) \\ \varphi_{X+a}(t) &= \exp(ita)\varphi_X(t) \end{aligned}}$$

et on en déduit, si X est une variable d'espérance m et d'écart-type σ, en posant $U = (X - m)/\sigma$:

$$\varphi_{\frac{X-m}{\sigma}}(t) = \varphi_U(t) = \exp\left(-\frac{itm}{\sigma}\right)\varphi_X\left(\frac{t}{\sigma}\right)$$

$$\varphi_X(t) = \exp(itm)\, \varphi_U(\sigma t)$$

2.6.1.3 Convolution

La fonction caractéristique se prête bien aux additions de variables aléatoires indépendantes : la fonction caractéristique d'une somme de variables indépendantes est égale au produit de leurs fonctions caractéristiques :

$$\boxed{\varphi_{X+Y}(t) = \varphi_X(t)\, \varphi_Y(t)}$$

En effet :

$$\varphi_{X_1+X_2}(t) = E\,[\exp(it\,(X_1 + X_2))] = E\,[\exp(itX_1)\exp(itX_2)]$$

si X_1 et X_2 sont indépendantes, il en est de même pour $\exp(itX_1)$ et $\exp(itX_2)$ et l'espérance du produit est alors égal au produit des espérances. Notons au passage qu'il ne s'agit donc pas d'une condition nécessaire et suffisante d'indépendance.

2.6.1.4 Cas d'une distribution symétrique

Supposons la loi de X symétrique par rapport à l'origine. Alors la fonction caractéristique de X est réelle :

$$\varphi_X(-t) = \int_{\mathbb{R}} \exp(-itx)\, dP_X(x) = \int_{\mathbb{R}} \exp(itx)\, dP_X(-x)$$

La première intégrale vaut $\overline{\varphi_X(t)}$ et la deuxième est égale à $\varphi_X(t)$ à cause de la symétrie car $dP_X(x) = dP_X(-x)$.

2.6.1.5 Dérivées à l'origine et moments non centrés

Notons tout d'abord que $\varphi_X(0) = 1$ car $\varphi_X(0) = \int_{\mathbb{R}} dP_X(x)$ P_X est une mesure de masse totale égale à 1.

Si les dérivées existent jusqu'à l'ordre k, on a :

$$\boxed{\varphi_X^{(k)}(0) = i^k E(X^k)}$$

En effet, $\varphi_X^{(k)}(t) = \int_{\mathbb{R}} (ix)^k \exp(itx)\, dP_X(x)$ par dérivation sous le signe somme. En particulier :

$$\varphi_X'(0) = iE(X)$$
$$\varphi_X''(0) = -E(X^2)$$

Si $\varphi_X(t)$ est indéfiniment dérivable, la formule de Mac-Laurin donne :

$$\varphi_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} i^k E(X^k)$$

2.6.1.6 Unicité et inversion de la fonction caractéristique

D'après les propriétés des transformées de Fourier, deux variables ayant même fonction caractéristique ont même loi de probabilité : la fonction caractéristique détermine donc de manière unique une distribution de probabilité d'où son nom.

Les formules d'inversion de la transformée de Fourier permettent d'obtenir la loi de X connaissant $\varphi_X(t)$.

THÉORÈME

Si $\int_{\mathbb{R}} |\varphi_X(t)|\, dt < \infty$ alors X admet une densité $f(x)$ continue et :

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \varphi_X(t) \exp(-itx)\, dt$$

Sinon, on a toujours le résultat suivant (admis) :

$$F(b) - F(a) = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{+T} \varphi_X(t) \frac{\exp(-ita) - \exp(-itb)}{it}\, dt$$

Une fonction quelconque n'est pas nécessairement une fonction de répartition ; de même, pour qu'une fonction $\varphi(t)$ soit une fonction caractéristique elle doit vérifier certaines propriétés. Le théorème suivant, que nous ne démontrerons pas, identifie les fonctions caractéristiques aux fonctions de « type positif ».

THÉORÈME (BOCHNER)

Pour qu'une fonction continue φ soit une fonction caractéristique, il faut et il suffit que pour toute famille finie $t_1, t_2, \ldots, t_n$ de réels et pour toute famille finie de complexes $z_1, z_2, \ldots, z_n$ on ait :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \varphi(t_i - t_j) z_i \bar{z}_j \geq 0$$

2.6.2 Fonctions caractéristiques des lois usuelles

2.6.2.1 *Lois discrètes*

- Loi de Bernoulli : $\varphi_X(t) = p \exp(it) + q$ avec $q = 1 - p$.
- Loi binomiale : Comme X est une somme de n variables de Bernoulli indépendantes, on trouve :

$$\varphi_X(t) = (p \exp(it) + q)^n$$

- Loi de Poisson :

$$\varphi_X(t) = \exp(\lambda (\exp(it) - 1))$$

En effet : $$E[\exp(itX)] = \sum_{x=0}^{\infty} \exp(itx) \exp(-\lambda) \frac{\lambda^x}{x!} = \exp(-\lambda) \sum_{x=0}^{\infty} \left(\frac{\lambda \exp(it)^x}{x!} \right)$$

$$= \exp(-\lambda) \exp(\lambda \exp(it))$$

2.6.2.2 *Lois continues*

- Loi uniforme sur [−a, a] :

$$\varphi_X(t) = \frac{\sin at}{at}$$

En effet : $$E[\exp(itX)] = \frac{1}{2a} \int_{-a}^{+a} \exp(itx)\, \mathrm{d}x = \frac{1}{2ait} [\exp(iat) - \exp(-iat)]$$

d'où le résultat avec $\exp(iat) = \cos at + i \sin at$.

- Lois gamma : Si X suit une loi γ_1, c'est-à-dire une loi exponentielle de paramètre 1, on a :

$$\varphi_{\gamma_1}(t) = \frac{1}{1 - it}$$

En effet : $$\varphi_{\gamma_1}(t) = \int_0^{\infty} \exp(itx) \exp(-x)\, \mathrm{d}x = \int_0^{\infty} \exp(-(1 - it)x)\, \mathrm{d}x$$

D'où, pour tout n entier :

$$\varphi_{\gamma_n}(t) = \frac{1}{(1 - it)^n}$$

car une γ_n est une somme de n γ_1 indépendantes.

Pour r quelconque, cette formule se généralise et $\varphi_{\gamma_r}(t) = \dfrac{1}{(1 - it)^r}$.

Remarquons que le calcul formel suivant conduit au résultat :

$$\int_0^{\infty} \exp(itx) \frac{1}{\Gamma(r)} \exp(-x)\, x^{r-1}\, \mathrm{d}x = \frac{1}{\Gamma(r)} \int_0^{\infty} \exp(-(1 - it)\, x)\, x^{r-1}\, \mathrm{d}x$$

en posant $(1 - it)x = u$:

$$= \frac{1}{\Gamma(r)}\int_0^{\infty} \exp(-u)u^{r-1} \frac{1}{(1 - it)^r} \mathrm{d}u = \frac{\Gamma(r)}{\Gamma(r)(1 - it)^r} = \frac{1}{(1 - it)^r}$$

Il convient cependant de justifier ce résultat car il s'agit d'une intégrale dans le champ complexe. Nous le laisserons au soin du lecteur.

- Loi de Laplace-Gauss : Si U est la loi LG(0 ; 1) :

$$\boxed{\varphi_u(t) = \exp(-t^2/2)}$$

On peut obtenir ce résultat directement car on sait que $E(U^k) = 0$ si k est impair et $E(U^{2k}) = \frac{(2k)!}{2^k\, k!}$.

D'après la formule de Mac-Laurin :

$$\varphi_u(t) = \sum_{k=0}^{\infty} \frac{t^{2k}}{2k!}(-1)^k \frac{2k!}{2^k k!}$$

$$= \sum_{k=0}^{\infty} \frac{\left(-\frac{t^2}{2}\right)^k}{k!} = \exp(-t^{2/2})$$

Remarquons qu'ici aussi un calcul formel (qui devrait être justifié par une intégration dans le plan complexe) donne le même résultat :

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \exp(itx)\, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}[x - it]^2 t^2/2\right) \mathrm{d}x$$

$$= \exp(-t^2/2) \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}[x - it]\right)^2 \mathrm{d}x$$

et l'intégrale vaut 1 car c'est l'intégrale de la densité d'une variable de Gauss imaginaire (!) de moyenne it et de variance 1.

Si X est une LG(m ; σ) :

$$\varphi_X(t) = \exp(itm) \exp\left(-\frac{t^2\sigma^2}{2}\right)$$

on en déduit que la somme de deux variables de Gauss indépendantes est encore une variable de Gauss :

$$X_1\, \mathrm{LG}(m_1 ; \sigma_1) \qquad X_2 \mathrm{LG}(m_2 ; \sigma_2)$$

$$\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) = \exp(it(m_1 + m_2)) \exp\left(-t^2\left(\frac{\sigma_1^2}{2} + \frac{\sigma_2^2}{2}\right)\right)$$

donc $X_1 + X_2$ suit une LG($m_1 + m_2$; $\sqrt{\sigma_1^2 + \sigma_2^2}$).

2.6.3 Fonctions génératrices

Il en existe deux formes assez voisines ; elles servent essentiellement à calculer les moments de variables aléatoires et de sommes de variables indépendantes car la fonction génératrice d'un produit de variables indépendantes est égale au produit de leurs fonctions génératrices.

- Pour des variables à valeurs entières positives, on utilisera la forme suivante :

$$g_X(t) = E(t^X) = \sum_{n \geq 0} t^n P(X = n)$$

Par dérivations successives en zéro, on trouve facilement que $g_X^{(n)}(0) = n!\, P(X = n)$, ce qui prouve que la fonction génératrice détermine la loi de probabilité de X.

Sous réserve d'existence, les dérivées successives en 1 sont égales aux moments factoriels :

$$g'_X(1) = E(X)$$
$$g''_X(1) = E(X(X-1))$$
$$g_X^{(n)}(1) = E(X(X-1)(X-2)\ldots(X-n+1))$$

- Pour des variables quelconques, on appelle fonction génératrice des moments :

$$M_X(t) = E(e^{tX})$$

qui est donc la transformée de Laplace de $-X$. Sous réserve d'existence, on a :

$$E(X^n) = M_X^{(n)}(0)$$

Les fonctions génératrices sont liées à la fonction caractéristique par :

$$g_X(t) = \varphi_X(-i \ln(t))$$
$$M_X(t) = \varphi_X(-it)$$

2.7 CONVERGENCES DES SUITES DE VARIABLES ALÉATOIRES

2.7.1 Les différents types de convergence

Une suite (X_n) de variables aléatoires étant une suite de fonctions de Ω dans $\mathbb{R}$, il existe diverses façons de définir la convergence de (X_n) dont certaines jouent un grand rôle en calcul des probabilités.

2.7.1.1 La convergence en probabilité

DÉFINITION

La suite (X_n) converge en probabilité vers la constante a si, $\forall \varepsilon$ et η (arbitrairement petits), il existe n_0 tel que $n > n_0$ entraîne :

$$P(|X_n - a| > \varepsilon) < \eta$$

On note alors $(X_n) \xrightarrow{P} a$.

On définit alors la convergence en probabilité vers une variable aléatoire X comme la convergence vers 0 de la suite $X_n - X$.

Lorsque $E(X_n) \to a$, il suffit de montrer que $V(X_n) \to 0$ pour établir la convergence en probabilité de X_n vers a. En effet, d'après l'inégalité de Bienaymé-Tchebycheff :

$$P(|X_n - E(X_n)| > \varepsilon) < \frac{V(X_n)}{\varepsilon^2}$$

On en déduit donc sans difficulté que $X_n - E(X_n) \xrightarrow{P} 0$, ce qui établit le résultat.

2.7.1.2 La convergence presque sûre ou convergence forte

Définissons d'abord l'égalité presque sûre de deux variables aléatoires :

DÉFINITION

X et Y sont égales presque sûrement si $P(\{\omega | X(\omega) \neq Y(\omega)\}) = 0$.

C'est l'égalité presque partout des fonctions mesurables. On définit donc ainsi des classes de variables aléatoires presque sûrement égales.

La convergence presque sûre se définit alors par :

DÉFINITION

La suite (X_n) converge presque sûrement vers X si :

$$P(\{\omega \mid \lim_{n \to \infty} X_n(\omega) \neq X(\omega)\}) = 0$$

et on note $X_n \xrightarrow{ps} X$.

En d'autres termes, l'ensemble des points de divergence est de probabilité nulle. Remarquons que la limite de (X_n) n'est pas unique mais que deux limites sont presque sûrement égales.

Il est immédiat de montrer que la convergence presque sûre implique la convergence en probabilité.

2.7.1.3 La convergence en moyenne d'ordre p

Si $E[(X_n - X)^p]$ existe, on a :

DÉFINITION

$(X_n) \to X$ *en moyenne d'ordre* p *si* $E[|X_n - X|^p] \to 0$.

La plus utilisée est la convergence en moyenne quadratique si $p = 2$.

La convergence en moyenne d'ordre p implique la convergence en probabilité.

2.7.1.4 La convergence en loi

Bien que la plus faible, elle est très utilisée en pratique car elle permet d'approximer la fonction de répartition de X_n par celle de X.

DÉFINITION

La suite (X_n) converge en loi vers la variable X de fonction de répartition F si, en tout point de continuité de F, la suite (F_n) des fonctions de répartition des X_n converge vers F. On note $X_n \xrightarrow{\mathcal{L}} X$.

Un théorème dû à Polya établit que si F est continue alors la convergence est uniforme.

Pour des variables discrètes, la convergence en loi vers une variable discrète s'exprime par $P(X_n = x) \rightarrow P(X = x)$.

C'est ainsi qu'on a établi la convergence de la loi binomiale vers la loi de Poisson.

Une suite de variables discrètes peut cependant converger en loi vers une variable continue (voir plus loin).

On montre également que, si (X_n) est une suite de variables de densités f_n et X une variable de densité f, alors :

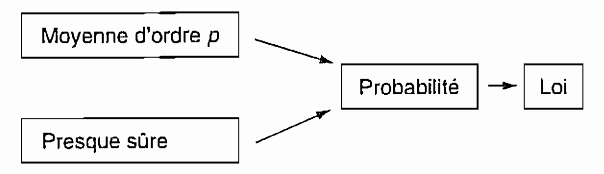
$$X_n \xrightarrow{\mathcal{L}} X \Rightarrow f_n(x) \rightarrow f(x) \qquad \forall x$$

La convergence en loi est intimement liée à la convergence des fonctions caractéristiques comme le précise le résultat fondamental suivant, que nous énoncerons sans démonstration :

THÉORÈME (LEVY-CRAMER-DUGUÉ)

Si $X_n \xrightarrow{\mathcal{L}} X$ alors $\varphi_{X_n}(t) \rightarrow \varphi_X(t)$ uniformément dans tout intervalle fini $[-u, u]$. Si la suite des fonctions caractéristiques $\varphi_{X_n}(t)$ converge vers une fonction φ dont la partie réelle est continue à l'origine, alors φ est une fonction caractéristique et la suite X_n converge en loi vers une variable aléatoire X dont φ est la fonction caractéristique.

La convergence en probabilité entraîne la convergence en loi et on a, pour résumer, la hiérarchie suivante des convergences :



2.7.2 Convergence en loi de la binomiale vers la loi de Laplace-Gauss (théorème de De Moivre-Laplace)

THÉORÈME

X_n étant une suite de variables binomiales $\mathcal{B}(n\,;p)$, alors $\dfrac{X_n - np}{\sqrt{npq}} \xrightarrow{\mathcal{L}} \text{LG}(0\,;1)$ en notant $q = 1 - p$.

Démonstration : La fonction caractéristique de X_n vaut $(p\exp(it) + 1 - p)^n$ donc celle de $\dfrac{X_n - np}{\sqrt{npq}}$ vaut :

$$\varphi(t) = \left(p\exp\left(\frac{it}{\sqrt{npq}}\right) + 1 - p\right)^n \exp\left(-\frac{itnp}{\sqrt{npq}}\right)$$

$$\ln\varphi = n\ln\left(p\left(\exp\left(\frac{it}{\sqrt{npq}}\right) - 1\right)\right) - \frac{itnp}{\sqrt{npq}}$$

Développons au deuxième ordre l'exponentielle ; il vient :

$$\ln\varphi \simeq n\ln\left(1 + p\left(\frac{it}{\sqrt{npq}} - \frac{t^2}{2npq}\right)\right) - \frac{itnp}{\sqrt{npq}}$$

puis le logarithme :

$$\ln\varphi \simeq n\left[\frac{pit}{\sqrt{npq}} - \frac{pt^2}{2npq} + \frac{p^2t^2}{2npq}\right] - \frac{itnp}{\sqrt{npq}}$$

soit :

$$\ln\varphi \simeq -\frac{t^2}{2q} + \frac{pt^2}{2q} = \frac{t^2}{2q}(p-1) = -\frac{t^2}{2}$$

car $p = 1 - q$.

$\varphi(t) \longrightarrow \exp(-t^2/2)$ qui est la fonction caractéristique de la loi normale centrée-réduite.

Application : Lorsque n est assez grand, on peut donc approximer la loi binomiale par la loi de Gauss. On donne généralement comme condition np et $nq > 5$.

Il convient cependant d'effectuer ce que l'on appelle la correction de continuité : la convergence de la loi binomiale vers la loi de Gauss se traduit par le fait que les extrémités des bâtons du diagramme de la binomiale $\mathcal{B}(n\,;p)$ sont voisines de la courbe de densité de la loi LG $(np\,;\sqrt{npq})$.

On obtient donc une valeur approchée de $P(X = x)$ par la surface sous la courbe de densité comprise entre les droites d'abscisse $x - \dfrac{1}{2}$ et $x + \dfrac{1}{2}$ (fig. 2.28).

$$P(X = x) \simeq P\left(\frac{x - \frac{1}{2} - np}{\sqrt{npq}} < U < \frac{x + \frac{1}{2} - np}{\sqrt{npq}}\right)$$

On aura alors :

$$P(X \le x) \simeq P\left(U < \frac{x + \frac{1}{2} - np}{\sqrt{npq}}\right)$$

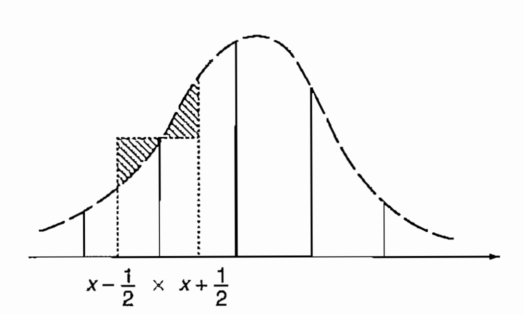


FIGURE 2.28

Exemple : X $\mathscr{B}(40\,;0.3)$ $np = 12$; $npq = 8.4$. La valeur exacte pour $P(X = 11)$ est 0.1319. La formule d'approximation avec une loi LG(12 ; $\sqrt{8.4}$) donne :

$$P\left(\frac{10.5-12}{\sqrt{8.4}} < U < \frac{11.5-12}{\sqrt{8.4}}\right)$$

soit : $P(-0.52 < U < -0.17) = P(0.17 < U < 0.52) = 0.6895 - 0.5675 = 0.122$

Soit une erreur de moins de 1 %.

Quant à $P(X \leq 11)$ qui vaut exactement 0.4406, l'approximation normale fournit $1 - P(U < 0.17)$ soit 0.4325. En l'absence de correction de continuité, on aurait trouvé $P\left(U < \frac{11-12}{\sqrt{8.4}}\right) = P(U < -0.35) = 1 - P(U < 0.35) = 0.3632$, ce qui est très imprécis.

2.7.3 Convergence de la loi de Poisson vers la loi de Gauss

THÉORÈME

Soit (X_λ) *une famille de variables* $\mathscr{P}(\lambda)$ alors si $\lambda \to \infty$, $\frac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{\mathscr{L}} \text{LG}(0\,;1)$.

Démonstration

$$\varphi_X(t) = \exp(\lambda)(\exp(it-1))$$

d'où :

$$\varphi_{\frac{X-\lambda}{\sqrt{\lambda}}}(t) = \exp\left(\lambda\left(\exp\left(\frac{it}{\sqrt{\lambda}}\right)-1\right)\right)\exp\left(-\frac{it\lambda}{\sqrt{\lambda}}\right)$$

$$= \exp\left(\lambda\exp\left(\frac{it}{\sqrt{\lambda}}\right) - \lambda - it\sqrt{\lambda}\right)$$

comme :
$$\exp\left(\frac{it}{\sqrt{\lambda}}\right) \simeq 1 + \frac{it}{\sqrt{\lambda}} - \frac{t^2}{2\lambda}$$

il vient :
$$\varphi_{\frac{X-\lambda}{\sqrt{\lambda}}}(t) \simeq \exp\left(\lambda + it\sqrt{\lambda} - \frac{t^2}{2} - \lambda - it\sqrt{\lambda}\right) = \exp\left(-\frac{t^2}{2}\right)$$

La figure 2.29 illustre l'approximation de la loi de Poisson $\mathscr{P}(\lambda)$ par la loi de Gauss de même espérance λ et de même écart-type $\sqrt{\lambda}$.

L'approximation est très satisfaisante pour $\lambda > 18$. On trouvera en annexe d'autres formules d'approximation plus précises. On a, ici encore, intérêt à effectuer la correction de continuité.

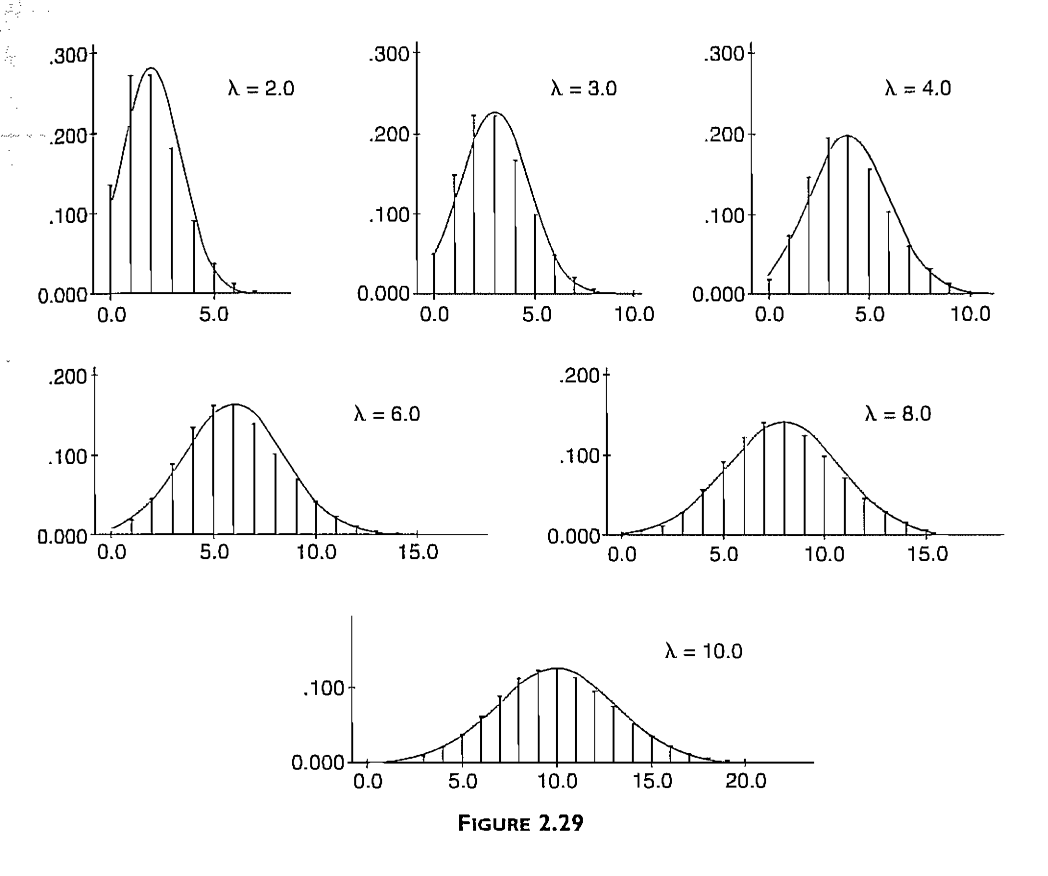


FIGURE 2.29

2.7.4 Le théorème central-limite

L'étude de sommes de variables indépendantes et de même loi joue un rôle capital en statistique.

Le théorème suivant connu sous le nom de ***théorème central-limite*** (il vaudrait mieux dire théorème de la limite centrée) établit la convergence vers la loi de Gauss sous des hypothèses peu contraignantes.

THÉORÈME

Soit (X_n) une suite de variables aléatoires indépendantes de même loi d'espérance μ et d'écart-type σ. Alors :

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma}\right) \xrightarrow{\mathcal{L}} \text{LG}(0\,;1).$$

Démonstration

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma}\right) = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}$$

Soit $\varphi_X(t)$ la fonction caractéristique de X ; la fonction caractéristique de $\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}$ est donc égale à $[\varphi_{\frac{X-\mu}{\sigma-\sqrt{n}}}(t)]^n$. Or $\frac{X - \mu}{\sigma\sqrt{n}}$ est une variable d'espérance nulle et de variance $1/n$.

Le développement en série de la fonction caractéristique de $\frac{X - \mu}{\sigma\sqrt{n}}$ commence par $1 - \frac{t^2}{2n}$, les termes suivants sont des infiniments petits d'ordre $1/n^2$.

Donc, en élevant à la puissance n, la fonction caractéristique de $\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}$ est équivalente à $\left(1 - \frac{t^2}{2n}\right)^n$ et tend si $n \longrightarrow \infty$ vers $\exp\left(-\frac{t^2}{2}\right)$ selon un résultat classique. ■

On remarque que, si les variables X_i sont des variables de Bernoulli, on retrouve comme cas particulier la convergence de la loi binomiale vers la loi de Gauss.

On peut démontrer un théorème encore plus général dû à Lindeberg :

THÉORÈME

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes pas forcément de même loi et d'espérance m_i et de variance σ_i^2. Soit $S_n^2 = \sum_{i=1}^{n} \sigma_i^2$ et $F_i(x)$ la fonction de répartition de $(X_i - m_i)$.

Si la condition suivante est réalisée :

$$\lim_{n\longrightarrow\infty}\left[\frac{1}{S_n^2}\sum_{i=1}^{n}\int_{|x|>\varepsilon S_n} x^2\, dF_i(x)\right] = 0$$

alors :

$$\frac{\sum_{i=1}^{n}(X_i - m_i)}{S_n} \xrightarrow{\mathcal{L}} U \in \text{LG}(0\,;1)$$

La condition de Lindeberg exprime que les variables $\frac{X_i - m_i}{S_n}$ sont « uniformément petites» avec une grande probabilité. Le résultat veut dire qu'à force d'ajouter de telles variables, on finit par obtenir une loi de Gauss.

Ce phénomène est souvent exprimé de la manière suivante : si une variable est la résultante d'un grand nombre de causes, petites, à effet additif, cette variable suit une loi de Gauss. On peut y voir la justification de l'emploi abondant et souvent abusif de la loi de Laplace-Gauss comme modèle.

Pour terminer, notons que l'existence des moments $E(X)$ et $V(X)$ est indispensable. La loi de Cauchy de densité $\frac{1}{\pi(1 + x^2)}$ sur $\mathbb{R}$ n'a aucun moment et fournit un contre-exemple classique : on montre que $\frac{X_1 + X_2 + \cdots + X_n}{n}$ a même loi que X quel que soit n.

3 Couples de variables aléatoires, conditionnement

L'étude de la loi de probabilité d'une variable aléatoire Y connaissant la valeur prise par une autre variable aléatoire X est fondamentale pour les problèmes d'approximation et de prévision. Il faut pour cela connaître en premier lieu la distribution de probabilité du couple (X, Y) qui est une application de $(\Omega, \mathscr{C}, P)$ dans $\mathbb{R}^2$ muni de sa tribu borélienne si il s'agit d'un couple de variables aléatoires réelles.

Il n'est cependant pas nécessaire que X et Y soient à valeurs dans $\mathbb{R}$.

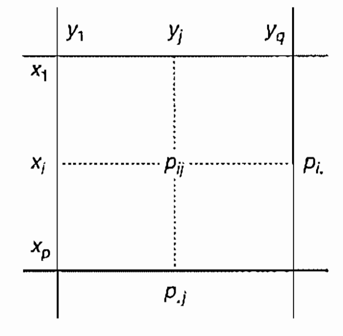
3.1 ÉTUDE D'UN COUPLE DE VARIABLES DISCRÈTES

On étudiera ici la distribution d'un couple de variables aléatoires à valeurs dans des ensembles finis ou dénombrables ; par exemple la distribution simultanée de la somme et du produit des points amenés par deux dés.

3.1.1 Lois associées à un couple (X, Y)

Supposons que X et Y prennent des valeurs x_i et y_j en nombre fini ou dénombrable.

3.1.1.1 Loi jointe



La loi du couple (X, Y) P_{XY} est alors entièrement définie par l'ensemble des nombres :

$$P_{XY}(x_i\,;\,y_j) = P(X = x_i \cap Y = y_j)$$

dans le cas fini cette loi de probabilité conjointe peut se mettre sous la forme d'une table.

On note $p_{ij} = P(X = x_i \cap Y = y_j)$ et bien sûr $\sum_i \sum_j p_{ij} = 1$.

3.1.1.2 Lois marginales

On appelle lois marginales les lois de probabilité de X et de Y pris séparément. On a d'après le théorème des probabilités totales :

– ***Loi marginale de X*** $\quad P(X = x_i) = \sum_{j=1}^{q} p_{ij} = p_{i.}$

– ***Loi marginale de Y*** $\quad P(Y = y_j) = \sum_{i=1}^{p} p_{ij} = p_{.j}$

3.1.1.3 Lois conditionnelles

Les évènements $\{X = x_i\}$ et $\{Y = y_j\}$ étant de probabilités non nulles on définit alors deux familles de lois conditionnelles selon que l'on connaît la « valeur » de X ou de Y. Rappelons qu'ici X et Y ne sont pas forcément des variables aléatoires réelles mais peuvent être des variables qualitatives. D'après le chapitre 1 on a :

– ***Lois conditionnelles de X si*** $Y = y_j$ **:**

$$P(X = x_i / Y = y_j) = \frac{p_{ij}}{p_{.j}} = \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}$$

– ***Lois conditionnelles de Y si*** $X = x_i$ **:**

$$P(Y = y_j / X = x_i) = \frac{p_{ij}}{p_{i.}} = \frac{P(X = x_i \cap Y = y_j)}{P(X = x_i)}$$

Le théorème des probabilités totales (deuxième forme) permet d'écrire :

$$P(X = x_i \cap Y = y_j) = \sum_{j=1}^{q} P(X = x_i / Y = y_j) P(Y = y_j)$$

$$= \sum_{i=1}^{p} P(Y = y_j / X = x_i) P(X = x_i)$$

Remarques :

- Pour deux événements B_1 et B_2 relatifs à Y et X on a :

$$P((Y \in B_2) \cap (X \in B_1)) = \sum_{x \in B_1} P(Y \in B_2 / X = x) P(X = x)$$

$$= \int_{B_1} P(Y \in B_2 / X = x) \, \mathrm{d}P_X(x)$$

formule qui servira pour étendre la notion de probabilité conditionnelle lorsque $X = x$ est de mesure nulle.

- Il arrive fréquemment dans les applications que l'on utilise la démarche inverse : on connaît la loi conditionnelle de Y à X fixé et celle de X et on en déduit alors la loi du couple.

Les formules de Bayes permettent d'exprimer une loi conditionnelle en fonction de l'autre :

$$P(X = x_i / Y = y_j) = \frac{P(Y = y_j / X = x_i) P(X = x_i)}{\sum_{i=1}^{p} P(Y = y_j / X = x_i) P(X = x_i)}$$

et :

$$P(Y = y_j / X = x_i) = \frac{P(X = x_i / Y = y_j) P(Y = y_j)}{\sum_{j=1}^{q} P(X = x_i / Y = y_j) P(Y = y_j)}$$

L'indépendance entre X et Y s'écrit :

$$p_{ij} = p_{i.}p_{.j} \quad \forall i \text{ et } j$$

ce qui revient à dire que les q lois conditionnelles de X à Y fixé (en faisant varier Y) sont identiques ; il en est de même pour les p lois conditionnelles de Y à X fixé.

3.1.2 Covariance et corrélation linéaire

La covariance a été introduite au chapitre 2 pour des variables numériques.

$$\text{cov}(X\,;Y) = E((X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y)$$

On a : $\text{cov}(X\,;X) = V(X)$ et $\text{cov}(Y\,;Y) = V(Y)$

On montrera plus loin que :

$$(\text{cov}(X\,;Y))^2 \leq V(X)V(Y)$$

ce qui permet de définir le coefficient de corrélation linéaire ρ, qui est donc toujours compris entre -1 et $+1$:

$$\rho = \frac{\text{cov}(X\,;Y)}{\sigma_X \sigma_Y}$$

Pour deux variables indépendantes $\rho = 0$. Cependant, la réciproque est en général inexacte et un coefficient de corrélation linéaire nul n'entraîne pas que les variables sont indépendantes. Deux exceptions notables où non-corrélation et indépendance sont équivalents : les couples $(X\,;Y)$ gaussiens (voir chapitre 4), et les couples de variables de Bernoulli (facile à montrer).

Les valeurs limites -1 et $+1$ sont atteintes si et seulement si il existe une relation linéaire entre Y et X.

3.1.3 Moments conditionnels

Supposons Y réelle mais pas nécessairement X qui peut être une variable qualitative. On peut alors définir, sous réserve de l'existence de ces expressions pour le cas dénombrable, l'espérance et la variance de Y à X fixé.

3.1.3.1 L'espérance conditionnelle

DÉFINITION

On appelle espérance de Y sachant que $X = x$ et on note $E(Y/X = x)$ la quantité définie par :

$$E(Y/X = x) = \sum_y yP(Y = y/X = x)$$

C'est donc l'espérance de Y prise par rapport à sa loi conditionnelle.

On note que $E(Y/X = x)$ est une fonction de x : $E(Y/X = x) = \varphi(x)$.

Cette fonction φ s'appelle ***fonction de régression***[1] ***de Y en X***. Son graphe est le lieu des moyennes conditionnelles de Y sachant X.

On voit donc que $E(Y/X = x)$ dépend des valeurs prises par X. On peut alors définir la variable aléatoire espérance conditionnelle, qui prend pour valeurs $E(Y/X = x)$ avec les probabilités $P(X = x)$:

DÉFINITION

On appelle variable aléatoire espérance conditionnelle de Y sachant X et on note $E(Y/X)$ la variable définie par :

$$E(Y/X) = \varphi(X)$$

Cette variable présente un certain nombre de propriétés remarquables.

Tout d'abord la linéarité comme conséquence de sa définition en tant qu'espérance :

$$E(Y_1 + Y_2/X) = E(Y_1/X) + E(Y_2/X)$$

mais surtout on a en prenant l'espérance de cette variable le :

THÉORÈME DE L'ESPÉRANCE TOTALE

$$\boxed{E[E(Y/X)] = E(Y)}$$

Démonstration

$$E[E(Y/X)] = \sum_x E(Y/X = x)P(X = x) = \sum_x \left(\sum_y yP(Y = y/X = x) \right) P(X = x)$$

$$= \sum_y y \sum_x P(Y = y/X = x)P(X = x) = \sum_y yP(Y = y) = E(Y)$$

Ce théorème est un outil très puissant pour calculer l'espérance mathématique d'une loi compliquée mais dont les lois conditionnelles sont simples : on voit même que l'on n'a pas besoin de connaître explicitement la loi de Y (voir plus loin).

Si $\psi(X)$ est une autre variable fonction de X on a $E[Y\psi(X)/X] = \psi(X)E[Y/X]$; la démonstration sans difficulté est omise. Concrètement cette formule signifie qu'à X fixé $\psi(X)$ est une constante et sort donc de l'espérance.

1 Ce terme de régression provient des travaux du statisticien Galton qui étudiait la taille des enfants Y en fonction de la taille de leur père X. Il avait constaté expérimentalement que la taille moyenne des fils dont le père avait une taille x supérieure à la moyenne $E(X)$ était elle-même supérieure à $E(Y)$ mais dans une moindre mesure $\frac{E(Y/X = x) - E(Y)}{x - E(X)}$ était inférieur à 1 ; il y avait donc régression au sens ordinaire du mot.

3.1.3.2 La variance conditionnelle

DÉFINITION

On appelle variance de Y sachant que $X = x$ et on note $V(Y/X = x)$ la quantité :

$$V(Y/X = x) = E[(Y - E(Y/X = x))^2/X = x]$$

Il s'agit donc de l'espérance conditionnelle du carré de l'écart à l'espérance conditionnelle.

Comme pour l'espérance, et puisque $V(Y/X = x) = \psi(x)$, on définit ensuite la variable aléatoire variance conditionnelle :

$$V(Y/X) = \psi(X) = E[(Y - E(Y/X))^2/X]$$

On a alors le résultat fondamental suivant :

THÉORÈME DE LA VARIANCE TOTALE

$$\boxed{V(Y) = E[V(Y/X)] + V[E(Y/X)]}$$

en donnant à l'espérance sa signification usuelle de moyenne on voit que la variance de Y est la somme de deux termes : la moyenne des diverses variances conditionnelles et la variance des diverses moyennes conditionnelles.

Démonstration

$$V(Y) = E[(Y - E(Y))^2] = E[(Y - E(Y/X) + E(Y/X) - E(Y))^2]$$

développons le carré en groupant $Y - E(Y/X)$ et $E(Y/X) - E(Y)$ il vient :

$$V(Y) = E[(Y - E(Y/X)^2] + 2E[(Y - E(Y/X))(E(Y/X) - E(Y))] + E[(E(Y/X) - E(Y))^2]$$

Le dernier terme est égal à $V[E(Y/X)]$ par définition de la variance puisque $E(Y)$ est l'espérance de $E(Y/X)$.

Le premier terme n'est autre que $E[V(Y/X)]$: en effet en appliquant le théorème de l'espérance totale :

$$E[(Y - E(Y/X))^2] = E[E[(Y - E(Y/X))^2/X]]$$

et on reconnaît l'expression de $V(Y/X)$. Notons que $V(Y/X)$ n'est pas égale à $(Y - E(Y/X))^2$ ce sont simplement deux variables ayant même espérance.

On vérifie que le double produit est nul en conditionnant à nouveau : l'espérance conditionnelle à X fixé de $(Y - E(Y/X))(E(Y/X) - E(Y))$ vaut alors :

$$[E(Y/X) - E(Y)][E(Y - E(Y/X))/X]$$

puisque $E(Y/X) - E(Y)$ est une constante à X fixé (voir la dernière propriété de l'espérance conditionnelle énoncée au sous-paragraphe précédent). Quant à :

$$E[(Y - E(Y/X))/X]$$

ce terme est nul, il suffit de développer. L'espérance conditionnelle du double produit est nul, il en est de même de son espérance.

(on trouvera plus loin une démonstration géométrique plus rapide et plus élégante) ■

3.1.3.3 Exemple d'utilisation de l'espérance et de la variance conditionnelle

Un examen se déroule sous forme d'un questionnaire à choix multiple (QCM) où on pose 20 questions ; chaque question comporte quatre réponses possibles, dont une et une seule est la bonne ; une réponse juste compte 1 point, sinon zéro.

On suppose que le programme de l'examen comporte 100 questions dont on tirera aléatoirement les 20 de l'examen.

Si l'on considère un candidat ayant appris une proportion p du programme, on étudie la distribution de sa note N.

Solution : Parmi les 20 questions, un certain nombre X va figurer dans la partie des 100 p questions révisées et fournir automatiquement X points. Les 20 questions étant tirées sans remise parmi les 100, la loi de X est une hypergéométrique $\mathcal{H}(100\,;\,20\,;\,p)$.

Un certain nombre de réponses pourront être devinées par le jeu du hasard parmi les $20 - X$ questions non révisées, soit Y ce nombre. A chaque question non révisée est associée une variable de Bernoulli de paramètre 1/4. Si $X = x$ est fixé, la loi de Y est alors une loi binomiale $\mathcal{B}(20 - x\,;\,1/4)$.

On a donc $N = X + Y$ avec $Y/X \ \mathcal{B}(20 - X\,;\,1/4)$. X et Y ne sont pas indépendantes puisque la distribution conditionnelle de $Y/X = x$ dépend de x.

Le calcul de la distribution de N conduit en tout état de cause à une expression difficilement manipulable :

$$P(N = n) = \sum_{x=0}^{x=n} P(X = x)P(Y = n - x/X = x)$$

$$= \sum_{x=0}^{x=n} \frac{C_{100p}^{x} C_{100(1-p)}^{20-x}}{C_{100}^{20}} C_{20-x}^{n-x} \left(\frac{1}{4}\right)^{n-x} \left(\frac{3}{4}\right)^{20-n}$$

On peut cependant trouver aisément $E(N)$ et $V(N)$:

- Calcul de $E(N)$:

$$E(N) = E(X) + E(Y) = E(X) + E[E(Y/X)]$$

$$E(X) = 20p \text{ (loi hypergéométrique)}$$

$$E(Y/X) = (20 - X)\frac{1}{4} = 5 - \frac{X}{4}$$

$$E[E(Y/X)] = 5 - \frac{E(X)}{4} = 5 - 5p$$

soit :

$$E(N) = 15p + 5$$

- Calcul de $V(N)$:

$$V(N) = E[V(N/X)] + V[E(N/X)]$$

$$V(N/X = x) = V[x + Y/X = x] = V[Y/X = x] = (20 - x)\frac{1}{4}\cdot\frac{3}{4}$$

$$V(N/X) = (20 - X)\frac{3}{16}\ E[V(N/X)] = 20(1 - p)\frac{3}{16} = \frac{15(1 - p)}{4}$$

$$E[N/X = x] = x + \frac{1}{4}(20 - x) = 5 + \frac{3x}{4}$$

$$E[N/X] = 5 + \frac{3X}{4}\ V[E(N/X)] = \frac{9}{16}V(X)$$

$$= \frac{9}{16}20p(1 - p)\frac{100 - 20}{100 - 1}$$

$$V[E(N/X)] = \frac{100p(1 - p)}{11}$$

$$V(N) = \frac{15(1 - p)}{4} + \frac{100p(1 - p)}{11} = (1 - p)\left[\frac{15}{4} + \frac{100p}{11}\right]$$

La figure 3.1 donne les variations de $E(N)$ et de $V(N)$ en fonction de p.

Un taux de révision de 0.6 à 0.7 devrait donc assurer la réussite à l'examen avec une forte probabilité.

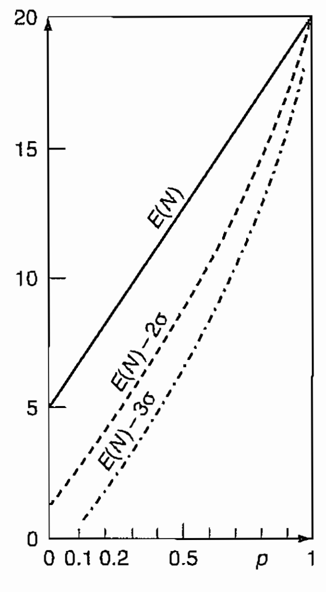


Figure 3.1

3.1.4 Extension au conditionnement d'une variable continue Y par une variable discrète X

Ce cas ne présente pas de difficulté. On définira d'abord la fonction de répartition conditionnelle :

$$P(Y < y/X = x) = \frac{P(Y < y \cap X = x)}{P(X = x)} = G(y/x)$$

puis si elle existe la densité conditionnelle $g(y/x)$ qui sera la dérivée de G en y.

La densité marginale de Y s'obtient par :

$$g(y) = \sum_x g(y/x)P(X = x)$$

Si $E(Y)$ existe, on prouve aisément que $E(Y/X = x)$ existe également et vaut :

$$E(Y/X = x) = \int_{\mathbb{R}} yg(y/x)\mathrm{d}y$$

Les formules de l'espérance totale et de la variance totale sont également valables. La formule de Bayes donne :

$$P(X = x/Y < y) = \frac{G(y/x)P(X = x)}{G(y)}$$

mais l'écriture formelle :

$$P(X = x/Y = y) = \frac{g(y/x)P(X = x)}{g(y)}$$

ne peut être pour l'instant justifiée car $P(Y = y) = 0$.

3.1.5 Somme d'un nombre aléatoire de variables *iid*

Le problème suivant est courant en assurance : au cours d'une période de temps donnée le nombre de sinistres survenus est une variable aléatoire N. Chaque sinistre a un coût aléatoire représenté par une variable X.

Le montant total des sinistres est alors :

$$S = X_1 + X_2 + \cdots X_N$$

Si les X_i sont indépendantes et de même loi, les théorèmes de l'espérance et de la variance totale, en conditionnant par N, permettent de montrer facilement que :

$$E(S) = E(N)E(X)$$
$$V(S) = E(N)V(X) + V(N)(E(X))^2$$

3.2 EXTENSION À DES VARIABLES QUELCONQUES

3.2.1 Lois conjointes et lois marginales d'un couple de variables aléatoires réelles

Si (X, Y) est à valeurs dans $\mathbb{R}^2$ rappelons que la fonction de répartition du couple $H(x, y)$ se définit par :

$$H(x, y) = P(X < x \cap Y < y)$$

Les fonctions de répartition marginales s'en déduisent immédiatement par :

$$F(x) = H(x\,;\infty) = P(X < x)$$
$$G(y) = H(\infty\,;y) = P(Y < y)$$

Si le couple (X, Y) admet une densité $h(x, y)$ on a :

$$h(x, y) = \frac{\partial^2 H}{\partial x \partial y}$$

les densités marginales s'obtiennent par :

$$f(x) = \int_{\mathbb{R}} h(x, y)\, \mathrm{d}y$$

$$g(x) = \int_{\mathbb{R}} h(x, y)\, \mathrm{d}x$$

Rappelons que si et seulement si les variables X et Y sont indépendantes on a :

$$H(x, y) = F(x)G(y) \quad \forall x\,;y$$
$$h(x, y) = f(x)g(y) \quad \forall x\,;y$$

3.2.2 Conditionnement

Le problème essentiel est de donner un sens aux expressions du type $P(Y \in B/X = x)$ et $E(Y/X = x)$ lorsque $X = x$ est un évènement de probabilité nulle ce qui est toujours le cas lorsque X est une variable admettant une densité.

3.2.2.1 *Présentation naïve*

Lorsque X est une variable continue on peut songer à définir la fonction de répartition conditionnelle de Y sachant que $X = x$ comme la limite pour ε tendant vers 0 de :

$$\frac{P(Y < y \cap (x < X < x + \varepsilon))}{P(x < X < x + \varepsilon)} = \frac{H(x + \varepsilon\,;y) - H(x\,;y)}{F(x + \varepsilon) - F(x)}$$

Lorsque X possède une densité $f(x)$ on « voit » que la limite de cette expression est $\dfrac{\partial H(x\,;y)}{\partial x}\Big/ f(x)$ et que si (X, Y) a une densité $h(x, y)$ la densité conditionnelle de Y à $X = x$ fixé vaut alors :

$$\frac{h(x\,;y)}{f(x)} = g(y/x)$$

On conçoit cependant aisément qu'une telle approche est peu rigoureuse et ne recouvre en plus qu'une partie du problème : dans certaines applications il faut pouvoir conditionner par rapport à une variable quelconque pas nécessairement à valeur dans $\mathbb{R}$ ni dans un ensemble fini. Pour définir une espérance conditionnelle il faut seulement que Y soit réelle et que $E(Y)$ existe.

3.2.2.2 Aperçus théoriques

Vu sa complexité nous ne donnerons que les résultats les plus importants sans rentrer dans les détails des démonstrations qui figurent dans les ouvrages de « Théorie des probabilités » (Neveu (1964) ou Métivier (1972) par exemple).

- Première présentation

X étant une variable aléatoire quelconque de $(\Omega, \mathcal{C}, P)$ dans un ensemble mesurable $(E, \mathcal{E})$ on définira la probabilité conditionnelle d'un événement A par rapport à X grâce au théorème suivant :

THÉORÈME

Soit $A \in \mathcal{C}$, alors $\forall B \in \mathcal{E}$ il existe une classe d'équivalence unique de fonctions de $(E, \mathcal{E})$ dans $[0\,;1]$ notée $P(A/X = x)$ telle que :

$$P(A \cap \{X \in B\}) = \int_B P(A/X = x)\, \mathrm{d}P_X(x)$$

La fonction $P(A/X = x)$ n'est pas unique car une modification de celle-ci sur un ensemble de probabilité P_X nulle ne change pas le résultat de l'intégrale.

Peut-on choisir un représentant de cette classe pour tout A qui définisse une loi de probabilité conditionnelle sur Ω ? Ce n'est pas sûr si X est quelconque et $P(./X = x)$ n'est pas nécessairement une mesure de probabilité : ici se trouve la difficulté majeure de la théorie. Si un tel choix est possible on dit que c'est une « version régulière » de la probabilité conditionnelle par rapport à X, notée $P(./X = x)$.

On peut alors définir l'espérance conditionnelle d'une variable Y intégrable par :

$$E(Y/X = x) = \int_\Omega Y(\omega)\, \mathrm{d}P(\omega/X = x)$$

- Deuxième présentation

Les ouvrages récents de théorie des probabilités préfèrent partir de la définition de l'espérance conditionnelle grâce au théorème suivant qui étend la formule de l'espérance totale en intégrant sur un événement quelconque de E au lieu d'intégrer sur E tout entier.

THÉORÈME

Soit Y une variable aléatoire réelle de $(\Omega, \mathscr{C}, P)$ dans $(\mathbb{R}, \mathscr{B})$ telle que $E(Y)$ soit fini, et X une variable quelconque de $(\Omega, \mathscr{C}, P)$ dans $(E, \mathscr{E})$ de loi de probabilité P_X.

Il existe alors une classe d'équivalence unique de fonctions P_X intégrables de $(E, \mathscr{E})$ dans $(\mathbb{R}, \mathscr{B})$ notée $E(Y/X = x)$ telle que :

$$\forall B \in \mathscr{E} \quad \int_{X^{-1}(B)} Y(\omega)\, \mathrm{d}P(\omega) = \int_B E(Y/X = x)\, \mathrm{d}P_X(x)$$

Ceci définit alors de manière (presque sûrement) unique la variable aléatoire espérance conditionnelle $E(Y/X)$.

On en déduit alors la probabilité d'un événement A quelconque de Ω conditionnellement à X en prenant pour Y la variable indicatrice de A :

$$P(A/X) = E(\mathbb{1}_A/X)$$

Comme $\mathbb{1}_A$ est intégrable la probabilité conditionnelle de A existe toujours. Le problème de l'existence d'une version régulière de la probabilité conditionnelle reste cependant entier, cette existence est nécessaire pour pouvoir calculer l'espérance conditionnelle par la formule :

$$E(Y/X = x) = \int_\Omega Y(\omega)\, \mathrm{d}P(\omega/X = x)$$

et pour pouvoir parler de distribution conditionnelle de Y sachant X.

La distribution conditionnelle de Y sachant $X = x$ est en effet définie comme la mesure image de $P(./X = x)$ par Y pour chaque x. Il faut donc que $P(./X = x)$ soit une mesure de probabilité sur Ω.

La preuve directe de l'existence de distributions conditionnelles dans les cas les plus usuels est donné par le théorème de Jirina : il suffit que E soit un espace métrique complet séparable (ou espace polonais), c'est-à-dire admettant un sous-ensemble partout dense, ce qui est le cas de $\mathbb{R}^p$.

3.2.2.3 *Ce qu'il faut retenir*

Il ressort des résultats précédents les propriétés utiles suivantes : si (X, Y) est un couple de variables aléatoires où Y est à valeurs dans $\mathbb{R}$ et X à valeurs dans un ensemble fini ou dénombrable, où à valeurs dans $\mathbb{R}$ ou $\mathbb{R}^p$:

- Il existe une mesure de probabilité conditionnelle $P(./X = x)$ sur Ω.
- Il existe une distribution conditionnelle de $Y/X = x$.

- Si $E(Y)$ existe, alors il existe une variable aléatoire espérance conditionnelle : $E(Y/X)$ qui prend les valeurs $E(Y/X = x)$ avec la loi de probabilité P_X :

$$E(Y/X = x) = \int_{\Omega} Y(\omega)\,\mathrm{d}P(\omega/X = x) = \int_{\mathbb{R}} y\,\mathrm{d}P(y/X = x)$$

et $E[E(Y/X)] = E(Y)$.
- Si $V(Y)$ existe on a $V(Y) = E(V(Y/X)) + V(E(X/Y))$.
- Si le couple (X, Y) est à valeur dans $\mathbb{R}^2$ et possède une densité $h(x, y)$ les densités conditionnelles existent et sont données par :

$$g(y/x) = \frac{h(x\,;y)}{f(x)} \qquad f(x/y) = \frac{h(x\,;y)}{g(y)}$$

et on a $E(Y/X = x) = \int_{\mathbb{R}} yg(y/x)\,\mathrm{d}y$ ainsi que les formules de Bayes pour les densités :

$$g(y/x) = \frac{f(x/y)g(y)}{\int_{\mathbb{R}} f(x/y)g(y)\,\mathrm{d}y} \qquad f(x/y) = \frac{g(y/x)f(x)}{\int_{\mathbb{R}} g(y/x)f(x)\,\mathrm{d}x}$$

- Lorsque l'une des variables est discrète et l'autre possède une densité il suffit de remplacer là où c'est nécessaire les intégrales par des sommes finies et les densités par des probabilités ponctuelles.

3.3 SYNTHÈSE GÉOMÉTRIQUE

Le cas où on n'étudie que des variables aléatoires réelles de moment d'ordre 2 fini est un des plus importants en pratique et est susceptible d'interprétations géométriques très éclairantes.

3.3.1 Espace de Hilbert des classes de variables aléatoires de carré intégrables

L'ensemble de toutes les variables aléatoires définies sur un même univers (en fait l'ensemble des classes de variables aléatoires presque partout égales) forme un espace de Hilbert L^2 si l'on le munit du produit scalaire :

$$\langle X, Y \rangle = E(XY) \qquad \text{et de la norme : } \|X\| = (E(X^2))^{1/2}$$

L'écart-type est donc la norme des variables centrées, et la covariance le produit scalaire des variables centrées.

Si l'on considère l'ensemble des variables aléatoires constantes, on obtient une droite D de L^2. Car si X est constante, aX l'est aussi.

L'espérance mathématique de X est alors la projection orthogonale de X sur cette droite (fig. 3.2) : en effet, on sait que le minimum de $E((X - a)^2)$ est atteint pour $a = E(X)$, ce qui définit la projection orthogonale de X sur D.

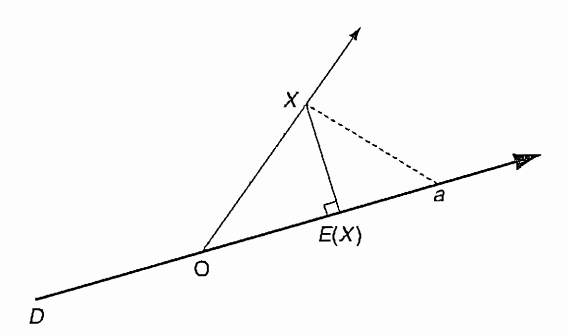


FIGURE 3.2

La formule de König-Huyghens :

$$E((X - a)^2) = V(X) + (E(X) - a)^2$$

s'interprète comme le théorème de Pythagore appliqué au triangle rectangle X, $E(X)$, a.

$E(X)$ est, en d'autres termes, la meilleure approximation de la variable X par une constante (au sens de la norme de L^2).

Comme $\text{cov}(X, Y) = \langle X - E(X) ; Y - E(Y) \rangle$ l'inégalité de Schwarz donne :

$$|\text{cov}(X, Y)| \leq \|X - E(X)\| \; \|Y - E(Y)\|$$

soit :

$$|\text{cov}(X, Y)| \leq \sigma_X \sigma_Y$$

Le cosinus de l'angle formé par $X - E(X)$ et $Y - E(Y)$ vaut donc $\dfrac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$. On retrouve le coefficient de corrélation linéaire ρ entre X et Y.

Dans cet espace, la non corrélation se traduit donc par l'orthogonalité

$\rho = \pm 1$ si $|\text{cov}(X, Y)| = \sigma_X \sigma_Y$ donc si $(X - E(X))$ et $(Y - E(Y))$ sont proportionnelles soit : $X - E(X) = a(Y - E(Y))$.

Le coefficient de corrélation linéaire est donc égal à ± 1 s'il y a une relation linéaire entre les deux variables X et Y.

La nullité de ce coefficient exclut la relation linéaire, mais n'exclut pas l'existence d'autres relations. Il est facile de fabriquer des contre-exemples de dépendance fonctionnelle avec un coefficient de corrélation linéaire nul : ainsi, X et X^2 ou $\sin X$ et $\cos X$ lorsque la loi de X est symétrique.

3.3.2 Espérance conditionnelle et projection

Soit L^2_X le sous-espace de L^2 constitué des variables aléatoires fonctions seulement de X du type $\varphi(X)$: L^2_X est convexe et contient la droite des constantes D.

C'est donc un sous-espace de Hilbert fermé.

Alors l'espérance conditionnelle de Y sachant X, $E(Y/X)$, s'interprète comme la projection orthogonale de Y sur L^2_X.

Soit en effet l'opérateur qui associe à toute variable aléatoire son espérance conditionnelle à X. C'est un opérateur linéaire ; pour montrer que c'est un projecteur orthogonal il suffit de vérifier qu'il est idempotent et auto-adjoint :

- il est idempotent : $E(E(Y/X)/X) = E(Y/X)$;
- et auto-adjoint : $\langle Z ; E(Y/X) \rangle = \langle E(Z/X) ; Y \rangle$.

En effet, les deux membres de cette relation sont égaux à $E[E(Z/X)E(Y/X)]$.

Le théorème de l'espérance totale $E(Y) = E(E(Y/X))$ est alors un cas particulier du théorème des trois perpendiculaires, comme l'illustre la figure 3.3.

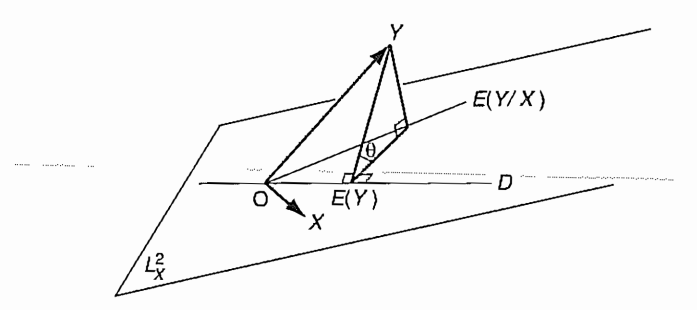


FIGURE 3.3

$E(Y/X)$ étant une projection orthogonale, ceci montre que le minimum de :

$$E[(Y - \varphi(X))^2]$$

est atteint pour $\varphi(X) = E(Y/X)$, résultat qui sera repris lors de l'étude de la régression. On peut dire que si $E(Y)$ est la meilleure approximation de Y par une constante, $E(Y/X)$ est la meilleure approximation de Y par une fonction de X.

Il est alors immédiat que le « résidu » $Y - E(Y/X)$ est non corrélé avec X par suite de l'orthogonalité.

Le théorème de la variance totale s'interprète comme le théorème de Pythagore appliqué au triangle rectangle Y, $E(Y)$, $E(Y/X)$:

$$\begin{aligned}\|Y - E(Y)\|^2 = V(Y) &= \|E(Y/X) - E(Y)\|^2 + \|Y - E(Y/X)\|^2 \\ &= V(E(Y/X)) + E[(Y - E(Y/X))^2] \\ &= V(E(Y/X)) + E[E(Y - E(Y/X))^2] \\ &= V(E(Y/X)) + E(V(Y/X))\end{aligned}$$

3.3.3 Rapport de corrélation de Y en X

Le coefficient de corrélation linéaire ρ est une mesure symétrique de dépendance, qui est maximale dans le cas de la liaison linéaire.

Le théorème de la variance totale permet de définir une autre mesure de liaison non symétrique cette fois : le rapport de corrélation $\eta_{Y/X}$ tel que :

$$\boxed{\eta^2_{Y/X} = \frac{V(E(Y/X))}{V(Y)}}$$

Ce rapport est le cosinus carré de l'angle formé par $Y - E(Y)$ et l'espace L^2_X.

On a donc : $$0 \leq \eta^2_{Y/X} \leq 1$$

PROPRIÉTÉ

Si $\eta^2_{Y/X} = 1$, $E(V(Y/X)) = 0$.

On en déduit donc que $V(Y/X) = 0$ presque sûrement, car c'est une variable positive. Ce qui veut dire qu'à X fixé la variance de Y est nulle, donc que Y ne prend qu'une seule valeur.

$$\eta^2_{Y/X} = 1 \Rightarrow Y = \varphi(X)$$

Le rapport de corrélation est maximal si Y est lié fonctionnellement à X.

PROPRIÉTÉ

Si $\eta^2_{Y/X} = 0$, $V(E(Y/X)) = 0$, $E(Y/X)$ est donc presque sûrement une constante.

On dit que Y est non corrélé avec X, il y a absence de dépendance en moyenne. C'est en particulier le cas si X et Y sont indépendantes mais la réciproque est inexacte. On montre en fait que l'indépendance entre Y et X est équivalente à l'orthogonalité des espaces L^2_X et L^2_Y engendrés par X et Y le long de la droite des constantes (fig. 3.4) :

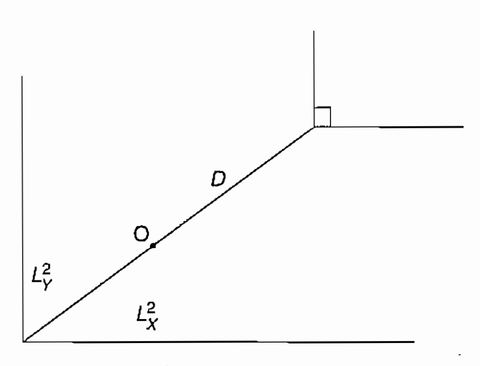


FIGURE 3.4 Indépendance de X et Y.

$\eta^2_{Y/X} = 0$ signifie seulement que $Y - E(Y)$ est orthogonal à L^2_X ;
η^2 est une mesure de liaison fonctionnelle alors que ρ est une mesure de liaison linéaire ;
$\eta^2_{Y/X}$ est toujours supérieur ou égal à ρ^2 car ρ^2 est le cosinus carré de l'angle formé par $Y - E(Y)$ avec le sous-espace de dimension 2 de L^2_X engendré par la droite des constantes D et la variable X.

Le cas ou $\eta^2_{Y/X} = \rho^2$ signifie donc que $E(Y/X)$ appartient à ce sous-espace de dimension 2, donc que :

$$E(Y/X) = \alpha + \beta X$$

c'est celui de la régression linéaire dont l'étude sera effectuée en détail au chapitre 16.

Si $E(Y/X) = \alpha + \beta X$, on ne peut trouver de transformation de X augmentant ρ.

En effet d'une part $\eta^2_{Y/X} = \sup_{\varphi} \rho^2(Y ; \varphi(X))$, et d'autre part la linéarité de la régression implique $\eta^2_{Y/X} = \rho^2(Y ; X)$.

Lorsque $(Y ; X)$ est un couple gaussien on a simultanément $E(Y/X) = \alpha + \beta X$ *et* $E(X/Y) = \gamma + \delta Y$

On en déduit le théorème suivant :

THÉORÈME

Si $(Y ; X)$ est un couple gaussien, on ne peut pas trouver de transformations $\varphi(X)$ et $\psi(Y)$ augmentant en valeur absolue le coefficient de corrélation :

$$\rho^2(\varphi(X) ; \psi(Y)) \leq \rho^2$$

Les prévisions optimales (en moyenne quadratique) sont donc linéaires.

4 Vecteurs aléatoires, formes quadratiques et lois associées

Ce chapitre présente les résultats les plus utiles pour l'étude des variables à plusieurs dimensions. Certaines démonstrations purement techniques seront omises.

4.1 GÉNÉRALITÉS SUR LES VECTEURS ALÉATOIRES RÉELS

Un vecteur aléatoire **X** est une application de $(\Omega, \mathscr{C}, P)$ dans un espace vectoriel réel, en général $\mathbb{R}^p$ muni de sa tribu borélienne.

En pratique $\mathbb{R}^p$ est muni de sa base canonique et on identifiera **X** au p-uple de variables aléatoires formé par ses composantes sur cette base $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

4.1.1 Fonction de répartition et densité

4.1.1.1 Fonction de répartition

F est une application de $\mathbb{R}^p$ dans $\mathbb{R}$ définie par :

$$F(x_1, x_2, \ldots, x_p) = P(X_1 < x_1, \ldots, X_p < x_p)$$

dont les propriétés se déduisent aisément de celles vues pour les couples de vecteurs aléatoires.

4.1.1.2 Densité

f si elle existe est définie par :

$$\boxed{f(x_1, \ldots, x_p) = \frac{\partial^p F}{\partial x_1 \partial x_2 \ldots \partial x_p}}$$

4.1.1.3 Changement de variables dans une densité

Effectuons le changement de variables défini par :

$$Y_i = \varphi_i(X_1, X_2, \ldots, X_p)$$

Les fonctions φ_i étant telles que le passage de $(X_1, X_2, \ldots, X_p)$ à $(Y_1, Y_2, \ldots, Y_p)$ est biunivoque. Nous désignerons en abrégé par φ la transformation :

$$\mathbf{X} \xrightarrow{\varphi} \mathbf{Y} \qquad \mathbf{Y} = \varphi(\mathbf{X})$$

La densité du vecteur **Y** s'obtient alors par la formule :

$$\boxed{g(\mathbf{y}) = \frac{f[\varphi^{-1}(\mathbf{y})]}{|\det \mathbf{J}|}}$$

où det **J**, appelé jacobien de la transformation, est tel que :

$$\det \mathbf{J} = \begin{vmatrix} \dfrac{\partial y_1}{\partial x_1} & \cdots\cdots & \dfrac{\partial y_p}{\partial x_1} \\ \vdots & & \vdots \\ \dfrac{\partial y_1}{\partial x_p} & \cdots\cdots & \dfrac{\partial y_p}{\partial x_p} \end{vmatrix}$$

$$(\det \mathbf{J})^{-1} = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \cdots\cdots & \dfrac{\partial x_1}{\partial y_p} \\ \vdots & & \vdots \\ \dfrac{\partial x_p}{\partial y_1} & \cdots\cdots & \dfrac{\partial x_p}{\partial y_p} \end{vmatrix} = \det \mathbf{J}^{-1}$$

La démonstration de cette propriété figure dans tous les ouvrages consacrés à l'intégration (changement de variable dans les intégrales multiples).

Si la transformation φ est linéaire de matrice **A** constante, $\mathbf{Y} = \mathbf{AX}$ (**A** doit être régulière) on a $\det \mathbf{J} = |\mathbf{A}|$. En particulier si **A** est une transformation orthogonale le jacobien vaut 1.

4.1.2 Fonction caractéristique

Soit **a** un vecteur non aléatoire de composantes $(a_1, a_2, \ldots, a_p)$.

DÉFINITION

On appelle fonction caractéristique du vecteur aléatoire **X** *la fonction de l'argument vectoriel* **a** *définie par :*

$$\varphi_{\mathbf{x}}(\mathbf{a}) = E[\exp(i\mathbf{a}'\mathbf{X})] = E[\exp(i(a_1X_1 + a_2X_2 + \cdots + a_pX_p))]$$

THÉORÈME

Les composantes $X_1, X_2, \ldots, X_p$ de **X** *sont indépendantes si et seulement si la fonction caractéristique de* **X** *est égale au produit des fonctions caractéristiques de ses composantes :*

$$\varphi_{\mathbf{x}}(\mathbf{a}) = \prod_{i=1}^{p} \varphi_{X_i}(a_i)$$

Si les X_i sont indépendantes l'espérance d'un produit de fonctions des X_i est égale au produit des espérances donc :

$$E[\exp(i\mathbf{a}'\mathbf{X})] = E[\exp(ia_1X_1)]\, E[\exp(ia_2X_2)] \ldots E[\exp(ia_pX_p)]$$

ce qui démontre une partie de la proposition.

La réciproque plus délicate utilise l'inversion de la fonction caractéristique et est omise.

Le résultat suivant fondamental permet de définir des lois de probabilités à p-dimensions à partir des lois unidimensionnelles.

THÉORÈME DE CRAMER-WOLD

La loi de $\mathbf{X}$ *est entièrement déterminée par celles de toutes les combinaisons linéaires de ses composantes.*

Posons en effet $Y = \mathbf{a}'\mathbf{X} = \sum_{i=1}^{p} a_iX_i$ et cherchons la fonction caractéristique de Y :

$$\varphi_Y(t) = E[\exp(itY)] = E[\exp(it\mathbf{a}'\mathbf{X})]$$

d'où $\varphi_Y(1) = \varphi_X(\mathbf{a})$. Si la loi de Y est connue pour tout $\mathbf{a}$ on connaît donc la fonction caractéristique de $\mathbf{X}$ donc la loi de $\mathbf{X}$.

4.1.3 Espérance et matrice de variance-covariance

Si μ_i désigne $E(X_i)$, on appelle par définition espérance de $\mathbf{X} = (X_1, \ldots, X_p)$ le vecteur certain :

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_p \end{bmatrix}$$

La matrice de variance-covariance $\boldsymbol{\Sigma}$ de $\mathbf{X}$ est définie par :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \mathrm{cov}(X_1, X_2) \ldots\ldots & \mathrm{cov}(X_1, X_p) \\ \cdot & \sigma_2^2 & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \sigma_p^2 \end{bmatrix} = E[\mathbf{X}\mathbf{X}'] - \boldsymbol{\mu}\boldsymbol{\mu}'$$

c'est une matrice carrée symétrique d'ordre p.

Si les variables X_i sont réduites, Σ s'identifie avec la matrice de corrélation :

$$\begin{bmatrix} 1 & \rho_{12} & \cdots\cdots & \rho_{1p} \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}$$

4.1.4 Transformations linéaires

Effectuons un changement de variable linéaire $\mathbf{Y} = \mathbf{AX}$ où $\mathbf{A}$ est une matrice quelconque de constantes (pas nécessairement carrée), alors :

$$\boldsymbol{\mu}_Y = \mathbf{A}\boldsymbol{\mu}_X$$
$$\boldsymbol{\Sigma}_Y = \mathbf{A}\boldsymbol{\Sigma}_X\mathbf{A}'$$

ce qui se démontre en appliquant les définitions.

En particulier si $\mathbf{A}$ est une matrice uniligne, Y est alors une variable aléatoire unidimensionnelle. Si $\mathbf{a}'$ désigne cette ligne $Y = \sum_{i=1}^{p} a_i X_i$ et $V(Y) = \mathbf{a}'\Sigma\mathbf{a}$. On a donc pour tout $\mathbf{a}$, $\mathbf{a}'\Sigma\mathbf{a} \geq 0$ car une variance est non négative. On en déduit le résultat suivant :

THÉORÈME

Une condition nécessaire et suffisante pour qu'une matrice Σ symétrique soit la matrice de variance d'un vecteur aléatoire est que Σ soit une matrice positive.

La réciproque s'établit à partir de la propriété classique suivante des matrices symétriques positives :

Toute matrice symétrique positive Σ peut s'écrire sous la forme $\Sigma = \mathbf{TT}'$ où $\mathbf{T}$ est définie à une transformation orthogonale près (si $\mathbf{T}$ convient, $\mathbf{S} = \mathbf{TU}$, où $\mathbf{U}$ est orthogonale, convient aussi ; une solution particulière est fournie par $\mathbf{T} = \Sigma^{1/2} = \mathbf{P}\Lambda^{1/2}\mathbf{P}'$ où $\mathbf{P}$ est la matrice des vecteurs propres normés de $\mathbf{T}$ et Λ la matrice diagonale des valeurs propres). Il suffit donc de partir d'un vecteur aléatoire $\mathbf{X}$ de matrice de variance $\mathbf{I}$, (par exemple un p-uple de variables indépendantes centrées-réduites) et de faire la transformation $\mathbf{Y} = \mathbf{TX}$ pour obtenir un vecteur aléatoire de matrice de variance Σ.

Si Σ est régulière, c'est-à-dire si les composantes de $\mathbf{X}$ ne sont pas linéairement dépendantes on peut trouver une transformation inverse qui « normalise » le vecteur $\mathbf{X}$.

THÉORÈME

Si Σ est régulière il existe une infinité de transformations linéaires $\mathbf{A}$, telles que $\mathbf{Y} = \mathbf{AX}$ soit un vecteur de matrice de variance $\mathbf{I}$.

Il suffit de prendre $\mathbf{A} = \mathbf{T}^{-1}$. Un choix particulièrement intéressant est celui de $\mathbf{T} = \Sigma^{1/2}$.

DÉFINITION

On appelle transformation de Mahalanobis la transformation définie par $\boldsymbol{\Sigma}^{-1/2}$.

$\mathbf{Y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ est alors un vecteur aléatoire centré-réduit à composantes non corrélées.

On en déduit aisément le résultat suivant :

THÉORÈME

La variable aléatoire $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = D^2$ *a pour espérance p.*

En effet $D^2 = \sum_{i=1}^{p} Y_i^2$ ou les Y_i sont d'espérance nulle et de variance 1. D est appelée distance de Mahalanobis de $\mathbf{X}$ à $\boldsymbol{\mu}$.

4.2 VECTEURS ALÉATOIRES GAUSSIENS : LA LOI MULTINORMALE

4.2.1 Définitions et fonction caractéristique

DÉFINITION

$\mathbf{X}$ *est un vecteur gaussien à p dimensions si toute combinaison linéaire de ses composantes* $\mathbf{a}'\mathbf{X}$ *suit une loi de Laplace-Gauss à une dimension.*

Le théorème de Cramer-Wold permet d'établir que la loi de $\mathbf{X}$ est ainsi parfaitement déterminée. On remarquera que la normalité de chaque composante ne suffit nullement à définir un vecteur gaussien.

La fonction caractéristique de $\mathbf{X}$ s'en déduit aisément (on supposera ici que $\mathbf{X}$ est centré ce qui ne nuit pas à la généralité).

THÉORÈME

$$\varphi_{\mathbf{X}}(\mathbf{a}) = \exp\left(-\frac{1}{2}\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}\right)$$ *où* $\boldsymbol{\Sigma}$ *est la matrice de variance de* $\mathbf{X}$.

En effet d'après le théorème de Cramer-Wold :

$$\varphi_{\mathbf{X}}(\mathbf{a}) = \varphi_Y(1) \qquad \text{où } Y = \mathbf{a}'\mathbf{X}$$

La loi de Y est par définition une gaussienne centrée de variance $V(Y) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ et la fonction caractéristique de Y est $\varphi_Y(t) = \exp\left(-\frac{t^2}{2}V(Y)\right)$ ce qui établit le résultat.

On en déduit le résultat fondamental suivant :

THÉORÈME

Les composantes d'un vecteur gaussien $\mathbf{X}$ *sont indépendantes si et seulement si* $\boldsymbol{\Sigma}$ *est diagonale, c'est-à-dire si elles sont non corrélées.*

On a en effet, si Σ est diagonale de termes σ_i^2 :

$$\varphi_{\mathbf{X}}(\mathbf{a}) = \exp\left(-\frac{1}{2}\sum_{i=1}^{p} a_i^2\sigma_i^2\right) = \prod_{i=1}^{p}\varphi_{X_i}(a_i)$$

On notera $N_p(\boldsymbol{\mu}\ ;\ \Sigma)$ la loi normale à p dimensions d'espérance $\boldsymbol{\mu}$ et de matrice de variance Σ.

4.2.2 Densité de la loi normale à p dimensions

Celle-ci n'existe que lorsque Σ est régulière.

THÉORÈME

Si Σ est régulière **X** *admet pour densité :*

$$f(x_1, x_2, \ldots, x_p) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

En effet $\mathbf{Y} = \Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu})$ est alors un vecteur gaussien dont les composantes sont centrées-réduites et indépendantes. **Y** a pour densité :

$$g(\mathbf{y}) = \prod_{i=1}^{p} g(y_i) = \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_i^2\right) = \frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{p}y_i^2\right)$$

Il suffit alors d'appliquer la formule du changement de variable ; le jacobien **J** vaut ici $\det\Sigma^{1/2} = (\det\Sigma)^{1/2}$ ce qui établit le résultat.

Les surfaces d'isodensité sont donc les ellipsoïdes d'équation $(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c$.

4.2.3 Cas particulier de la loi normale à deux dimensions

Si l'on introduit ρ coefficient de corrélation linéaire entre X_1 et X_2 :

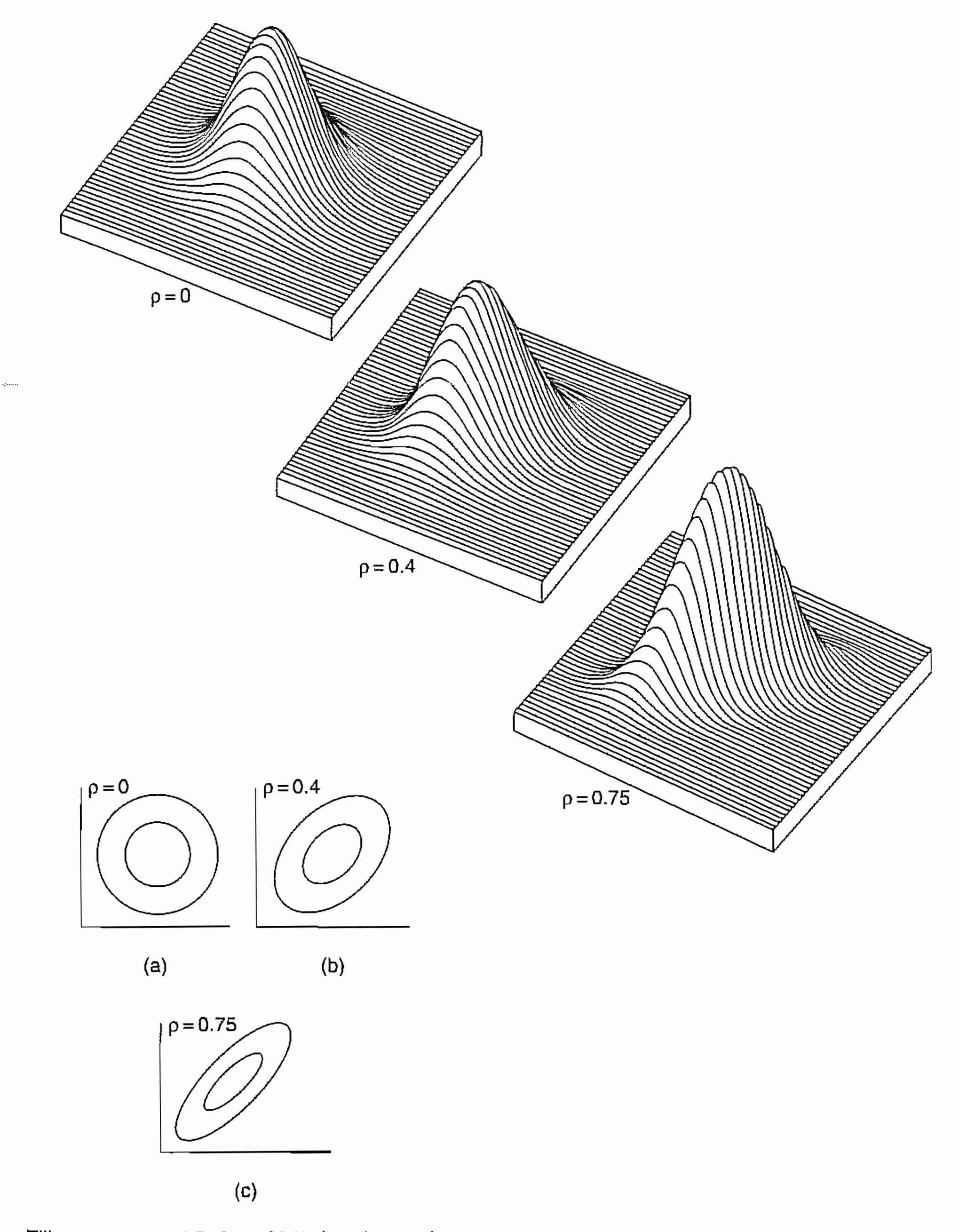
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

d'où : $$\det\Sigma = (\sigma_1\sigma_2)^2(1-\rho^2)$$

et : $$\Sigma^{-1} = \frac{1}{\det\Sigma}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-m_1}{\sigma_1}\right)^2 - 2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-m_2}{\sigma_2}\right)^2\right]\right\}$$

La figure 4.1 représente quelques surfaces de densité correspondant à $\sigma_1 = \sigma_2 = 1$ et à diverses valeurs de ρ ainsi que les ellipses d'isodensité dans le plan X_1, X_2.



Ellipses contenant 50 % et 90 % des observations

FIGURE 4.1 (d'après Bhattacharyya et Johnson, 1977).

4.2.4 Lois conditionnelles (sans démonstration)

Partitionnons $\mathbf{X}$ en deux sous-vecteurs $\mathbf{X}_1$ et $\mathbf{X}_2$ à k et $p-k$ composantes respectivement d'espérance $\mathbf{m}_1$ et $\mathbf{m}_2$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \hline \mathbf{X}_2 \end{bmatrix}$$

La matrice de variance-covariance se partitionne en 4 blocs :

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Si l'on cherche la loi du vecteur $\mathbf{X}_1$, conditionnée par $\mathbf{X}_2$ on a les résultats suivants :

THÉORÈME

La loi de $\mathbf{X}_1/\mathbf{X}_2$ est une loi multinormale à p dimensions :

- *d'espérance : $E[\mathbf{X}_1/\mathbf{X}_2] = \mathbf{m}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \mathbf{m}_2)$;*
- *de matrice variance-covariance : $\boldsymbol{\Sigma}_{11/2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.*

On constate donc que la régression de $\mathbf{X}_1$ en $\mathbf{X}_2$ est linéaire.

Les termes de $\boldsymbol{\Sigma}_{11/2}$ s'appellent les covariances partielles cov $(i, j \mid 2)$, desquelles on déduit les corrélations partielles :

$$\rho_{ij/2} = \frac{\text{cov}(i, j \mid 2)}{\sigma_{ii/2}\,\sigma_{jj/2}}$$

Les variances conditionnelles ne dépendent pas des valeurs prises par $\mathbf{X}_2$: il y a « homoscédasticité ».

4.2.5 Théorème central-limite multidimensionnel

De même que pour des lois à une dimension on peut établir le résultat suivant :

Soit $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ une suite de vecteurs aléatoires indépendants de même loi, d'espérance $\boldsymbol{\mu}$ et de matrice de variance $\boldsymbol{\Sigma}$ alors :

THÉORÈME

$$\boxed{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}) \xrightarrow{\mathcal{L}} N_p(\mathbf{0} ; \boldsymbol{\Sigma})}$$

4.3 FORMES QUADRATIQUES DÉFINIES SUR UN VECTEUR GAUSSIEN ET LOIS DÉRIVÉES

4.3.1 Lois du χ^2 (khi-deux)

DÉFINITION

$U_1, U_2, \ldots, U_p$ étant p variables LG(0 ; 1) *indépendantes, on appelle loi du khi-deux à p degrés de liberté* (χ_p^2) *la loi de la variable* $\sum_{i=1}^{p} U_i^2$.

C'est donc la loi de la somme des carrés des composantes d'un vecteur gaussien centré et de matrice de variance **I**.

On en déduit immédiatement que la somme de deux variables χ^2 indépendantes à p et q degrés de liberté est encore une variable χ^2, à $p + q$ degrés de liberté.

La loi du χ^2 se déduit de la loi γ par une simple transformation.

Prenons en effet un χ_1^2, c'est-à-dire le carré d'une variable de Gauss. D'après un résultat établi au chapitre 2, la densité de $T = U^2$ est :

$$g(t) = \frac{1}{\sqrt{2\pi}} t^{-1/2} \exp\left(-\frac{t}{2}\right)$$

Puisque $\sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)$ on en déduit que $\frac{U^2}{2} = \gamma_{1/2}$. On a donc la propriété suivante :

PROPRIÉTÉ

Si X est une variable γ_r, $2X$ est un χ_{2r}^2

On en déduit donc par transformation les propriétés de la loi du χ^2 :

$$E(\chi_p^2) = p \qquad V(\chi_p^2) = 2p$$

Densité : $g(\chi_p^2) = \dfrac{1}{2^{p/2}\Gamma\left(\frac{p}{2}\right)} \exp\left(\dfrac{-\chi^2}{2}\right)(\chi^2)^{p/2-1}$ (fig. 4.2).

A. *Fonction caractéristique*

Elle se déduit de celle de la loi γ :

$$\varphi_{\chi_p^2}(t) = \frac{1}{(1-2it)^{p/2}}$$

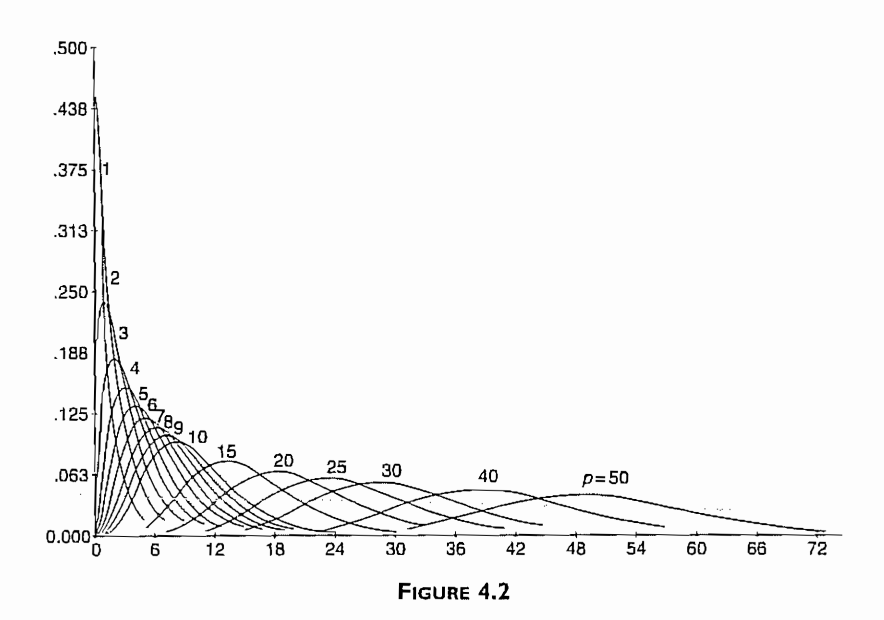


FIGURE 4.2

B. Approximation du χ^2 par la loi normale

Lorsque $p > 30$ on peut admettre que $\sqrt{2\chi^2} - \sqrt{2p-1}$ est distribué comme une LG(0 ; 1), soit :

$$\chi^2_p \simeq \frac{\left(U + \sqrt{2p-1}\right)^2}{2} \qquad \text{(approximation de Fisher)}$$

ou (mieux) que :

$$\left[\left(\frac{\chi^2_p}{p}\right)^{1/3} + \frac{2}{9p} - 1\right]\sqrt{\frac{9p}{2}} \simeq U$$

soit :

$$\chi^2_p \simeq p\left(U\sqrt{\frac{2}{9p}} + 1 - \frac{2}{9p}\right)^3 \qquad \text{(approximation de Wilson-Hilferty)}$$

Cette dernière approximation, très précise, est correcte même pour des valeurs faibles de p. On trouvera en annexe des formules exactes permettant de calculer la fonction de répartition du χ^2.

La table A1.6 donne les fractiles de la loi de χ^2 jusqu'à 100 degrés de liberté. On peut donc en déduire ceux de la loi γ_r pour des valeurs de r allant de $1/2$ à 50 par demi-entier.

4.3.2 Formes quadratiques

Sous certaines conditions, des formes quadratiques définies sur des vecteurs gaussiens suivent des lois du χ^2. Ces résultats sont fondamentaux en statistique dans les problèmes de décomposition de variance.

THÉORÈME

Si **X** *suit une loi normale à p dimensions d'espérance* $\boldsymbol{\mu}$ *et de matrice de variance* $\boldsymbol{\Sigma}$ *régulière alors :*

$$D^2 = (\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}) \quad \text{suit une loi du } \chi^2_p$$

Il suffit de se souvenir que $D^2 = \sum_{i=1}^{p} Y_i^2$ où les Y_i sont des LG(0 ; 1) indépendantes.

Considérons maintenant **Y** vecteur gaussien centré-réduit à composantes indépendantes et cherchons la loi d'une forme quadratique générale $Q = \mathbf{Y'AY} = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} Y_i Y_j$.

Nous allons établir la forme de la fonction caractéristique de Q ce qui permettra après de déduire dans quels cas Q suit une loi du χ^2.

THÉORÈME

$$\varphi_Q(t) = [\det(\mathbf{I} - 2it\mathbf{A})]^{-1/2}$$

Démonstration

$$\varphi_Q(t) = E[\exp(itQ)] = E[\exp(it\mathbf{Y'AY})]$$

Écrivons $\mathbf{A} = \mathbf{P}'\boldsymbol{\Lambda}\mathbf{P}$ où **P** est la matrice orthogonale des vecteurs propres et $\boldsymbol{\Lambda}$ la matrice diagonale des valeurs propres λ_i de **A** :

$$\mathbf{Y'AY} = \sum_{j=1}^{p} \lambda_j Z_j^2 \quad \text{en posant } \mathbf{Z} = \mathbf{PY}$$

P étant orthogonale **Z** est encore un vecteur gaussien centré-réduit à composantes indépendantes.

Donc : $$\varphi_Q(t) = E\left[\exp\left(it \sum_{i=1}^{p} \lambda_j z_j^2\right)\right] = \prod_{j=1}^{p} \varphi_{z_j^2}(\lambda_j t)$$

or Z_j^2 est un χ^2_1 d'où : $$\varphi_Q(t) = \prod_{j=1}^{p} (1 - 2i\lambda_j t)^{-1/2}$$

or si λ_j est valeur propre de **A**, $1 - 2i\lambda_j t$ est valeur propre de $\mathbf{I} - 2it\mathbf{A}$, donc :

$$\prod_{j=1}^{p} (1 - 2i\lambda_j t) = \det(\mathbf{I} - 2it\mathbf{A})$$

On peut également donner la démonstration suivante plus directe mais utilisant des gaussiennes complexes.

Démonstration

$$E[\exp(it\mathbf{Y}'\mathbf{AY})] = \int_{\mathbb{R}^p} \exp(it\mathbf{y}'\mathbf{AY})g(\mathbf{y})\,d\mathbf{y} = \int_{\mathbb{R}^p} \exp(it\mathbf{y}'\mathbf{Ay})\frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{y}\right)d\mathbf{y}$$

$$= \frac{1}{(2\pi)^{p/2}}\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}\mathbf{y}'(\mathbf{I} - 2it\mathbf{A})\mathbf{y}\right)d\mathbf{y}$$

Or si l'on considère une loi gaussienne de matrice de variance $\mathbf{\Sigma} = (\mathbf{I} - 2it\mathbf{A})^{-1}$ on sait que :

$$\int_{\mathbb{R}^p} \frac{1}{(2\pi)^{p/2}(\det \mathbf{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{\Sigma}^{-1}\mathbf{y}\right)d\mathbf{y} = 1$$

d'où : $E[\exp(itQ)] = (\det \mathbf{\Sigma})^{1/2} = [\det(\mathbf{I} - 2it\mathbf{A})]^{-1/2}$ ■

On peut donc établir la propriété suivante :

Théorème

$Q = \mathbf{Y}'\mathbf{AY}$ suit une loi du χ^2 si et seulement si $\mathbf{A}$ est un projecteur orthogonal, c'est-à-dire si $\mathbf{A}^2 = \mathbf{A}$. Le rang de $\mathbf{A}$ est alors le degré de liberté du χ^2.

En effet si $\mathbf{A}^2 = \mathbf{A}$ $\lambda_j = 0$ ou 1 et $\varphi_Q(t)$ est la fonction caractéristique d'un χ^2_r. La réciproque est alors immédiate.

Considérons maintenant deux formes quadratiques Q_1 et Q_2 de matrice $\mathbf{A}_1$ et $\mathbf{A}_2$ définies sur $\mathbf{Y}$.

Théorème de Craig

Q_1 et Q_2 sont indépendantes si et seulement si $\mathbf{A}_1\,\mathbf{A}_2 = \mathbf{0}$.

Démonstration

$$\varphi_{Q_1Q_2}(t_1, t_2) = E[\exp(it_1Q_1 + it_2Q_2)] = [\det(\mathbf{I} - 2it_1\mathbf{A}_1 - 2it_2\mathbf{A}_2)]^{-1/2}$$

Comparons cette expression au produit des deux fonctions caractéristiques de Q_1 et Q_2.

$$\varphi_{Q_1}(t_1)\varphi_{Q_2}(t_2) = [\det(\mathbf{I} - 2it_1\mathbf{A}_1)\det(\mathbf{I} - 2it_2\mathbf{A}_2)]^{-1/2}$$
$$= [\det(\mathbf{I} - 2it_1\mathbf{A}_1 - 2it_2\mathbf{A}_2 - 4t_1t_2\mathbf{A}_1\mathbf{A}_2)]^{-1/2}$$

on aura $\varphi_{Q_1}(t_1)\varphi_{Q_2}(t_2) = \varphi_{Q_1Q_2}(t_1t_2)\ \forall t_1t_2$ si et seulement si $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$ ce qui établit le théorème. ■

Nous pouvons enfin énoncer le résultat le plus important concernant les formes quadratiques qui généralise la propriété d'additivité du χ^2 :

THÉORÈME DE COCHRAN

Soient $Q_1, Q_2, \ldots, Q_k$ k formes quadratiques sur **Y** *telles que* $\sum_{j=1}^{k} Q_j = \mathbf{Y'Y} = \sum_{i=1}^{p} Y_i^2$ *c'est-à-dire réalisant une décomposition du carré de norme de* **Y**.

Alors les trois conditions suivantes sont équivalentes :

- $\sum_{j=1}^{k} \text{rang}\,(Q_j) = p$;
- *chaque Q_j est une variable de χ^2 ;*
- *les Q_j sont indépendantes.*

Ce théorème n'est que la version probabiliste d'un théorème classique d'algèbre linéaire que voici.

Soit k matrices symétriques $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ d'ordre p telles que $\sum_{j=1}^{k} \mathbf{A}_j = \mathbf{I}_p$.

Alors les trois conditions suivantes sont équivalentes :

- $\sum_i \text{rang}\ \mathbf{A}_i = p$;
- $\mathbf{A}_i^2 = \mathbf{A} \qquad \forall i$;
- $\mathbf{A}_i\mathbf{A}_j = \mathbf{0} \qquad$ pour $i \neq j$.

La démonstration, sans difficulté, est laissée au soin du lecteur. Géométriquement ce théorème est une extension du théorème de Pythagore et de sa réciproque à la décomposition d'un vecteur et donc de son carré de norme, sur des sous-espaces deux à deux orthogonaux. L'orthogonalité est ici synonyme d'indépendance pour des vecteurs gaussiens.

4.3.3 Loi du F de Fisher-Snedecor

Cette loi, liée au rapport de deux formes quadratiques indépendantes joue un grand rôle en statistique (loi du rapport des variances de deux échantillons indépendants par exemple).

X et Y étant des variables suivant indépendamment des lois χ^2_n et χ^2_p, on définit :

$$\boxed{F(n\,;p) = \frac{X/n}{Y/p}}$$

La densité de F s'obtient aisément par transformation de celle d'une bêta II car $X/2$ et $Y/2$ suivent des lois $\gamma_{n/2}$ et $\gamma_{p/2}$:

$$g(f) = \frac{1}{B\left(\frac{n}{2}\,;\frac{n}{2}\right)} \frac{\left(\frac{n}{p}\right)^{n/2} f^{\,n/2-1}}{\left(1 + \frac{n}{p}f\right)^{(n+p)/2}}$$

$$E(F) = \frac{p}{p-2} \qquad \text{et} \qquad V(F) = 2\frac{p^2}{n}\frac{n+p-2}{(p-2)^2(p-4)}$$

Cette loi est tabulée en annexe ce qui permet d'obtenir les distributions des lois bêta I et bêta II ; on a en effet les relations suivantes :

- si Y suit une loi bêta II(n, p), alors $\frac{pY}{n}$ est un $F(2n, 2p)$;
- si X suit une loi bêta I(n, p), alors $\frac{p}{n}\frac{X}{1-X}$ est un $F(2n, 2p)$.

4.3.4 Loi de Student

Soit une variable aléatoire U suivant une LG(0, 1) et X une variable aléatoire suivant indépendamment de U une loi χ^2_n. On définit alors la variable de Student T_n à n degrés de liberté comme étant :

$$\boxed{T_n = \frac{U}{\sqrt{\frac{X}{n}}}}$$

On a :

$$E(T_n) = 0 \qquad \text{si } n > 1$$

$$V(T_n) = \frac{n}{n-2} \qquad \text{si } n > 2$$

$$\mu_3 = 0 \qquad \text{si } n > 3$$

$$\mu_4 = \frac{3n^2}{(n-2)(n-4)} \qquad \text{si } n > 4$$

$$\gamma_2 = 3 + \frac{6}{n-4} \qquad \text{si } n > 4$$

Pour $n = 1$ la loi de Student est la loi de Cauchy, loi du quotient de deux variables aléatoires de Laplace-Gauss indépendantes, dont la densité est :

$$f(t) = \frac{1}{\pi(1+t^2)}$$

Cette loi ne possède aucun moment fini. De manière générale la densité de T_n est :

$$f(t) = \frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}$$

si $n \to \infty$, $T_n \xrightarrow{\mathcal{L}}$ LG(0 ; 1), ainsi que l'expression des moments le laissait supposer.

On a la relation suivante entre les variables de Student et de Fisher-Snedecor :

$$(T_n)^2 = F(1\,;n)$$

La figure 4.3 donne les densités de T_n pour diverses valeurs du degré de liberté :

$$n = 1, 2, 5, 10, 50.$$

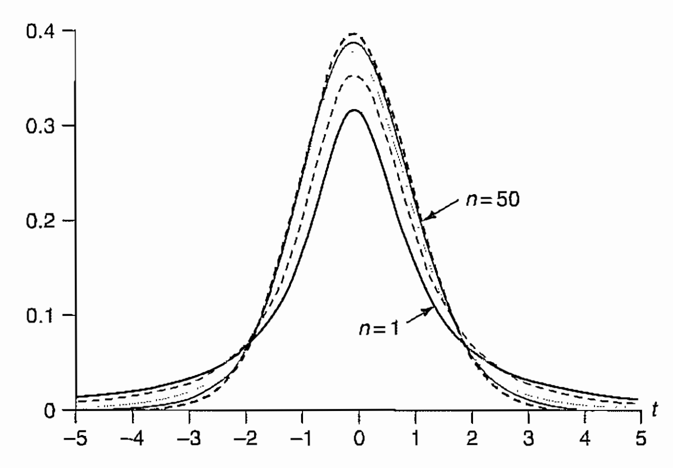


FIGURE 4.3 Densité de probabilité de la variable de Student

On remarquera le comportement particulier de la loi de Cauchy T_1. qui a des queues de distribution très importantes :

$$P(|T_1| > 2) = 0.29$$

4.4 LA LOI MULTINOMIALE, INTRODUCTION AU TEST DU χ^2

Comme son nom l'indique cette loi généralise la loi binomiale.

4.4.1 Le schéma de l'urne à *k* catégories

Considérons une partition de Ω en k événements de probabilité $p_1, p_2, \ldots, p_k$ (fig. 4.4).

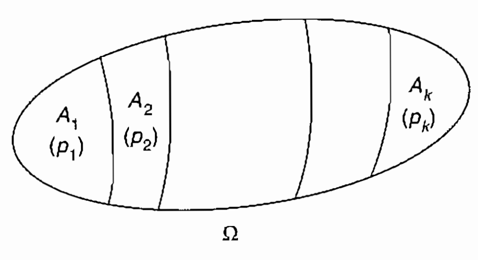


FIGURE 4.4

On répète alors indépendamment n fois l'expérience aléatoire et on compte les nombres de réalisations respectives des A_i : $N_1, N_2, \ldots, N_k$.

Le vecteur aléatoire (discret) $(N_1, N_2, \ldots, N_k)$ suit alors par définition une loi multinomiale d'effectif n et de paramètres $p_1, p_2, \ldots, p_k$.

Ce schéma se produit en particulier dans des problèmes de sondages : une population est partagée en k catégories et on tire **avec** remise n individus ; on compte ensuite les effectifs de cet échantillon appartenant aux diverses catégories.

On l'observe également lors du dénombrement des réalisations d'une variable aléatoire X :

L'ensemble des valeurs de X est partagé en k classes de probabilités p_i et on compte sur un ensemble de n individus les nombres d'individus appartenant à chacune de ces classes (fig. 4.5) : (c'est la démarche utilisée pour construire un histogramme, voir chapitre 5).

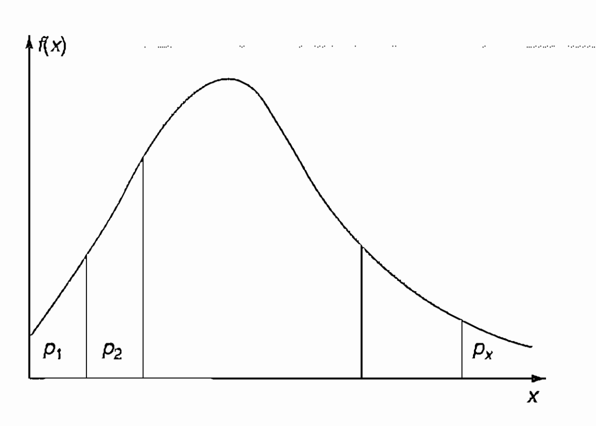


Figure 4.5

Par construction, les composantes $N_1, N_2, \ldots, N_k$ du vecteur multinomial sont linéairement dépendantes : $\sum_{i=1}^{k} N_i = n$ et on a bien sûr $\sum_{i=1}^{k} p_i = 1$.

Chaque composante N_i suit une loi binomiale $\mathscr{B}(n\,;p_i)$ donc $E(N_i) = np_i$ et $V(N_i) = np_i(1 - p_i)$.

La loi conditionnelle de N_i sachant $N_j = n_j$ est également une loi binomiale :

$$\boxed{\mathscr{B}\left(n - n_j\,;\ \frac{p_i}{1 - p_j}\right)}$$

Il suffit de remarquer que tout se passe comme si il restait à tirer $n - n_j$ individus dans une population à $k - 1$ catégories : la catégorie A_j étant éliminée la probabilité conditionnelle d'observer $A_i/\overline{A}_j$ vaut $\frac{p_i}{1 - p_j}$.

La loi du k-uple est alors donnée par :

$$P(N_1 = n_1 ; N_2 = n_2 ; N_k = n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

4.4.2 Espérance et matrice de variance

Comme chaque N_i suit une loi $\mathcal{B}(n ; p_i)$ on a :

$$\boldsymbol{\mu} = \begin{bmatrix} np_1 \\ np_2 \\ . \\ . \\ . \\ np_k \end{bmatrix}$$

Pour établir la covariance entre N_i et N_j il suffit de remarquer que le vecteur multinomial est une somme de n vecteurs indépendants de même loi que le vecteur $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ tel que $X_i = 0$ ou 1 avec les probabilités $1 - p_i$ et p_i ; un seul des X_i étant nul. Les X_i sont les indicatrices des catégories $A_1, A_2, \ldots, A_k$ pour un des n tirages.

On a alors $E(X_i X_j) = 0$ si $i \neq j$ d'où $\text{cov}(X_i, X_j) = -E(X_i)E(X_j) = -p_i p_j$.

La covariance d'une somme étant la somme des covariances on en déduit :

$$\text{cov}(N_i, N_j) = -np_i p_j \qquad \text{si } i \neq j$$

La matrice de variance-covariance de la loi multinomiale est donc :

$$n\boldsymbol{\Sigma} = n \begin{bmatrix} p_1(1 - p_1) & -p_1 p_2 & \ldots & -p_1 p_k \\ -p_1 p_2 & p_2(1 - p_2) & \ldots & -p_2 p_k \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ -p_1 p_k & -p_2 p_k & \ldots & p_k(1 - p_k) \end{bmatrix}$$

Cette matrice n'est pas régulière car $\sum_{i=1}^{k} N_i = n$ (on remarque que les sommes en lignes et en colonnes sont nulles).

4.4.3 Lois limites lorsque $n \to \infty$

D'après le théorème central limite multidimensionnel, comme $(N_1, N_2, \ldots, N_k)$ est une somme de n vecteurs aléatoires indépendants et de même loi, on a :

$$\frac{1}{\sqrt{n}}(N_1 - np_1 ; N_2 - np_2 ; \ldots ; N_k - np_k) \xrightarrow{\mathcal{L}} N_k(\mathbf{0} ; \boldsymbol{\Sigma})$$

La loi limite est dégénérée (elle n'admet pas de densité) car $\sum_{i=1}^{k}(N_i - np_i) = 0$.

Cependant si l'on supprime par exemple la dernière composante on a alors un vecteur limite gaussien non dégénéré et :

$$\mathbf{X} = \frac{1}{\sqrt{n}}(N_1 - np_1 ; N_2 - np_2 ; \ldots ; N_{k-1} - np_{k-1}) \xrightarrow{\mathcal{L}} N_{k-1}(\mathbf{0} ; \mathbf{\Sigma}^*)$$

où $\mathbf{\Sigma}^*$ s'obtient en supprimant la dernière ligne et la dernière colonne de $\mathbf{\Sigma}$.

Par une simple vérification on trouve :

$$(\mathbf{\Sigma}^*)^{-1} = \begin{bmatrix} \left(\frac{1}{p_1} + \frac{1}{p_k}\right) & \frac{1}{p_k} & \ldots & \frac{1}{p_k} \\ \frac{1}{p_k} & \left(\frac{1}{p_2} + \frac{1}{p_k}\right) & \ldots & \frac{1}{p_k} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \frac{1}{p_k} & \ldots\ldots\ldots\ldots & \ldots & \left(\frac{1}{p_{k-1}} + \frac{1}{p_k}\right) \end{bmatrix}$$

Appliquons alors le premier théorème sur les formes quadratiques :

$$D^2 = \mathbf{X}'(\mathbf{\Sigma}^*)^{-1}\mathbf{X} \rightarrow \chi^2_{k-1}$$

En développant on a :

$$\begin{aligned} D^2 &= \sum_{i=1}^{k-1}\frac{(N_i - np_i)^2}{np_i} + \sum_{i=1}^{k-1}\frac{(N_i - np_i)^2}{np_k} + \sum_{i \neq j}\sum\frac{(N_i - np_i)(N_j - np_0)}{np_k} \\ &= \sum_{i=1}^{k-1}\frac{(N_i - np_i)^2}{np_i} + \frac{1}{np_k}\left(\sum_{i=1}^{k-1}(N_i - np_i)\right)^2 \\ &= \sum_{i=1}^{k-1}\frac{(N_i - np_i)^2}{np_i} + \frac{1}{np_k}(N_k - np_k)^2 \end{aligned}$$

car $\sum_{i=1}^{k-1} N_i = n - N_k$ et $\sum_{i=1}^{k-1} np_i = n - np_k$.

Il vient donc :

$$\boxed{\sum_{i=1}^{k}\frac{(N_i - np_i)^2}{np_i} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{k-1}}$$

Ce résultat capital est à la base du test du khi-deux permettant de comparer une distribution d'observations $N_1, N_2, \ldots, N_k$ à une distribution théorique de probabilités $p_1, p_2, \ldots, p_k$ (voir chapitre 14, paragraphe 14.6.2.1).

4.5 LOIS DE WISHART, DE HOTELLING, DE WILKS

Ces lois jouent un rôle essentiel en statistique mathématique multidimensionnelle.

4.5.1 Loi de Wishart

DÉFINITION

Une matrice **M** (p, p) *a une distribution de Wishart* $W_p(n\,;\boldsymbol{\Sigma})$ *si* **M** *peut s'écrire* $\mathbf{M} = \mathbf{X}'\mathbf{X}$ *où* **X** *est une matrice* (n, p) *aléatoire définie de la façon suivante : les* n *lignes de* **X** *sont des vecteurs aléatoires gaussiens de même loi* $N_p(\mathbf{0}\,;\boldsymbol{\Sigma})$ *indépendants.*

X représente donc un échantillon de n observations indépendantes d'une loi normale multidimensionnelle.

Nous allons voir que cette loi généralise d'une certaine façon la loi du χ^2. Si $p = 1$ on a en effet :

$$W_1(n\,;\sigma^2) = \sigma^2\chi^2_n = \sum_{i=1}^{n} x_i^2$$

On montre que la densité de la loi de Wishart est :

$$f(\mathbf{M}) = \frac{|\mathbf{M}|^{(n-p-1)/2}\exp\left(-\frac{1}{2}\operatorname{Trace}\boldsymbol{\Sigma}^{-1}\mathbf{M}\right)}{2^{np/2}\pi^{p(p-1)/4}|\boldsymbol{\Sigma}|^{n/2}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+i-1)\right)}$$

avec $\mathbf{M} > 0$ pour la mesure de Lebesgue dans $\mathbb{R}^{p(p+1)/2}$ (en effet **M** doit être symétrique et semi définie positive).

On rapprochera cette formule de celle de la densité d'un χ^2.

On note également que la fonction caractéristique de la loi de Wishart $W_p(n\,;\,\Sigma)$ est :

$$E[\exp(i\mathbf{TM})] = |\mathbf{I} - i\mathbf{T}\Sigma|^{-n/2}$$

où **T** est une matrice (p, p).

Rappelons que la fonction caractéristique d'un χ^2_n est $\varphi_{\chi^2_n}(t) = (1 - 2it)^{-n/2}$.

On a : $E(\mathbf{M}) = n\Sigma$ et $E(\mathbf{M}^{-1}) = \dfrac{1}{n-p-1}\Sigma^{-1}$ si $n - p - 1 > 0$

Pour tout vecteur constant **a** :

$$\frac{\mathbf{a}'\mathbf{Ma}}{\mathbf{a}'\Sigma\mathbf{a}} \quad \text{suit une loi } \chi^2_n$$

En effet on vérifie sans peine que $\mathbf{a}'\mathbf{Ma}$ est une matrice de Wishart $W_1(n\,;\,\mathbf{a}'\Sigma\mathbf{a})$ car $\mathbf{a}'\mathbf{Ma} = \mathbf{a}'\mathbf{X}'\mathbf{Xa}$ où $\mathbf{Xa}$ suit $N_1(\mathbf{0}\,;\,\mathbf{a}'\Sigma\mathbf{a})$.

On peut montrer également, mais la démonstration est délicate, que $\dfrac{\mathbf{a}'\Sigma^{-1}\mathbf{a}}{\mathbf{a}'\mathbf{M}^{-1}\mathbf{a}}$ suit une loi χ^2_{n-p+1}.

Ces deux propriétés se généralisent avec des vecteurs aléatoires.

PROPRIÉTÉ

Soit **x** un vecteur aléatoire (de loi quelconque) indépendant de **M** alors :

$$\frac{\mathbf{x}'\mathbf{Mx}}{\mathbf{x}'\Sigma\mathbf{x}} \quad \text{et} \quad \frac{\mathbf{x}'\Sigma^{-1}\mathbf{x}}{\mathbf{x}'\mathbf{M}^{-1}\mathbf{x}}$$

suivent les lois χ^2_n et χ^2_{n-p+1} respectivement et sont des variables indépendantes de **x**.

$\mathbf{a}'\mathbf{Ma}$ et $\mathbf{b}'\mathbf{Mb}$ sont indépendantes si $\mathbf{a}'\Sigma\mathbf{b} = 0$.

4.5.2 La loi du T^2 de Hotelling

Cette distribution généralise celle de Student (ou plutôt son carré). C'est celle d'une variable unidimensionnelle.

DÉFINITION

Soit **x** *un vecteur aléatoire normal* $N_p(\mathbf{0}\,;\,\mathbf{I})$ *et* **M** *une matrice de Wishart* $W_p(n\,;\,\mathbf{I})$, *indépendante de* **x** *; alors la quantité* $n\mathbf{x}'\mathbf{M}^{-1}\mathbf{x}$ *suit par définition une loi du* T^2 *de Hotelling de paramètres p et n.*

Par abus de notation, on posera : $T^2_p(n) = n\mathbf{x}'\mathbf{M}^{-1}\mathbf{x}$

PROPRIÉTÉ

Si $\mathbf{x}$ suit une loi $N_p(\boldsymbol{\mu} ; \boldsymbol{\Sigma})$ et $\mathbf{M}$ une loi de Wishart indépendante de $\mathbf{x}$ $W_p(n ; \boldsymbol{\Sigma})$ alors $n(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ suit une loi $T_p^2(n)$.

La démonstration évidente utilise la transformation de Mahalanobis $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ et le fait que $\boldsymbol{\Sigma}^{-1/2}\mathbf{M}\boldsymbol{\Sigma}^{-1/2}$ est une $W_p(n ; \mathbf{I}_p)$.

$n\mathbf{x}'\mathbf{M}^{-1}\mathbf{x}$ suit ce qu'on appelle une loi de Hotelling décentrée $T_p^2(n, \lambda^2)$ où $\lambda^2 = \boldsymbol{\mu}'\boldsymbol{\Sigma}\boldsymbol{\mu}$ est le paramètre de décentrement.

La loi du T^2 de Hotelling s'identifie à celle de Fisher-Snedecor selon la formule :

$$\boxed{T_p^2(n) = \frac{np}{n - p + 1} F(p ; n - p + 1)}$$

En effet, on peut écrire avec $\mathbf{x}$ $N_p(\mathbf{0} ; \mathbf{I})$:

$$T_p^2(n) = \frac{n\mathbf{x}'\mathbf{M}^{-1}\mathbf{x}}{\mathbf{x}'\mathbf{x}}\mathbf{x}'\mathbf{x}$$

où $\dfrac{\mathbf{x}'\mathbf{x}}{\mathbf{x}'\mathbf{M}^{-1}\mathbf{x}}$ est un χ^2_{n-p+1} indépendant de $\mathbf{x}$ donc de $\mathbf{x}'\mathbf{x}$ qui est un χ^2_p d'où :

$$T_p^2(n) = n\frac{\chi^2_p}{\chi^2_{n-p+1}}$$

On voit que pour $p = 1$, $T_1^2(n) = F(1 ; n)$ c'est-à-dire le carré de la variable de Student à n degrés de liberté.

Notons que : $$E(T_p^2(n)) = \frac{np}{n - p - 1}$$

4.5.3 La loi du lambda (Λ) de Wilks

Cette loi joue un grand rôle en analyse de variance multidimensionnelle où elle généralise celle de Fisher-Snedecor : elle concerne les rapports de variance généralisée qui sont des déterminants de matrices de Wishart. Λ est une variable unidimensionnelle.

DÉFINITION

Soit $\mathbf{A}$ *et* $\mathbf{B}$ *deux matrices de Wishart* $W_p(m ; \boldsymbol{\Sigma})$ *et* $W_p(n ; \boldsymbol{\Sigma})$ *indépendantes où* $m \geq p$, *alors le quotient :*

$$\frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|} = \frac{1}{|\mathbf{A}^{-1}\mathbf{B} + \mathbf{I}|} = \Lambda$$

a une distribution de Wilks de paramètres p, m *et* n, $\Lambda(p, m, n)$ *(cette distribution ne dépend pas de* $\boldsymbol{\Sigma}$*).*

$\mathbf{A}$ et $\mathbf{B}$ étant des matrices positives Λ est une variable comprise entre 0 et 1.

Remarquons que Λ s'exprime en fonction des valeurs propres θ_i de $\mathbf{A}^{-1}\mathbf{B}$:

$$\Lambda = \prod_{i=1}^{p}(1 + \theta_i)^{-1}$$

$\Lambda(p, m, n)$ et $\Lambda(n, m + n - p, p)$ ont la même distribution.

On peut se ramener à la loi de Fisher-Snedecor dans quatre cas simples :

$$\frac{1 - \Lambda(p, m, 1)}{\Lambda(p, m, 1)} = \frac{p}{m - p + 1} F(p\,;\, m - p + 1)$$

$$\frac{1 - \Lambda(1, m, n)}{\Lambda(1, m, n)} = \frac{n}{m} F(n\,;\, m)$$

$$\frac{1 - \sqrt{\Lambda(p, m, 2)}}{\sqrt{\Lambda(p, m, 2)}} = \frac{p}{m - p + 1} F(2p\,;\, 2(m - p + 1))$$

$$\frac{1 - \sqrt{\Lambda(2, m, r)}}{\sqrt{\Lambda(2, m, r)}} = \frac{n}{m - 1} F(2n\,;\, 2(m - 1))$$

Si m est grand on peut utiliser l'approximation de Bartlett :

$$-\left[m - \frac{1}{2}(p - n + 1)\right] \ln \Lambda(p, m, n) \simeq \chi^2_{np}$$

DEUXIÈME PARTIE

Statistique exploratoire

5 Description unidimensionnelle de données numériques

La plupart du temps les données se présentent sous la forme suivante : on a relevé sur n unités appelées « individus » p variables numériques. Lorsque n et p sont grands on cherche à synthétiser cette masse d'informations sous une forme exploitable et compréhensible. Une première étape consiste à décrire séparément les résultats obtenus pour chaque variable : c'est la description unidimensionnelle, phase indispensable, mais insuffisante (voir chapitre suivant), dans toute étude statistique.

On considérera donc ici qu'on ne s'intéresse qu'à une variable X, appelée encore caractère, dont on possède n valeurs $x_1, x_2, \ldots, x_n$.

La synthèse de ces données se fait sous forme de ***tableaux***, de ***graphiques*** et de ***résumés numériques***. C'est ce que l'on appelle couramment la « statistique descriptive » dout l'usage a été considérablement facilité par l'informatique.

5.1 TABLEAUX STATISTIQUES

Leur présentation diffère légèrement selon la nature des variables.

5.1.1 Variables discrètes ou qualitatives

Pour chaque valeur ou modalité x_i de la variable on note n_i le nombre d'occurrences (ou effectif) de x_i dans l'échantillon, $\Sigma n_i = n$, et f_i la fréquence correspondante $f_i = n_i/n$ (on utilise en fait le plus souvent le pourcentage $100f_i$.).

Le tableau statistique se présente en général sous la forme :

x_i	n_i	f_i

■ **Exemple 1 :** Le recensement général de la population française en 1999 donne la répartition des 23 810 161 ménages, selon la variable X nombre de personnes du ménage.

Rappelons qu'un ménage est composé de toutes les personnes habitant normalement dans un logement, quels que soient leurs liens de parenté. Les ménages sont donc ici les individus ou unités statistiques.

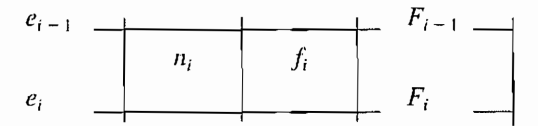
Tableau 5.1

Nombre de personnes	Pourcentage
1	31.0
2	31.1
3	16.2
4	13.8
5	5.5
6 et plus	2.4

5.1.2 Variables continues ou assimilées

On regroupe les valeurs en k classes d'extrémités $e_0, e_1, \ldots, e_k$ et l'on note pour chaque classe $[e_{i-1}, e_i[$ l'effectif n_i et la fréquence f_i ainsi que les fréquences cumulées $F_i = \sum_{j=1}^{i} f_j$, ou proportion des individus pour lesquels $X < e_i$.

Le tableau statistique se présente en général comme suit :



Par convention, la borne supérieure d'une classe est toujours exclue de cette classe.

Exemple 2 : Le magazine *Capital* a donné pour 100 villes françaises les valeurs du taux de la taxe d'habitation.

Tableau 5.2

Ville	Taux taxe d'habitation	Zone Géographique	Ville	Taux taxe d'habitation	Zone Géographique
Aix-en-Provence	18.94	Sud-Est	Aubervilliers	12.45	Ile-de-France
Ajaccio	22.06	Sud-Est	Aulnay-sous-Bois	15.59	Ile-de-France
Amiens	17.97	Nord	Avignon	22.41	Sud-Est
Angers	18.86	Ouest	Beauvais	15.37	Nord
Annecy	14.97	Sud-Est	Belfort	16.20	Est
Antibes	14.30	Sud-Est	Besançon	20.20	Est
Antony	11.07	Ile-de-France	Béziers	22.14	Sud-Ouest
Argenteuil	16.90	Ile-de-France	Blois	17.07	Centre
Arles	24.49	Sud-Est	Bordeaux	22.11	Sud-Ouest
Asnières-sur-Seine	10.13	Ile-de-France	Boulogne-Billancourt	9.46	Ile-de-France

Ville	Taux taxe d'habitation	Zone Géographique	Ville	Taux taxe d'habitation	Zone Géographique
Bourges	15.77	Centre	Maisons-Alfort	10.30	Ile-de-France
Brest	25.99	Ouest	Marseille	21.93	Sud-Est
Brive-la-Gaillarde	15.82	Centre	Mérignac	19.39	Sud-Ouest
Caen	16.12	Ouest	Metz	16.62	Est
Calais	23.36	Nord	Montauban	12.72	Sud-Ouest
Cannes	19.72	Sud-Est	Montpellier	21.40	Sud-Ouest
Chalon-sur-Saone	17.30	Centre	Montreuil	13.67	Ile-de-France
Chambéry	18.71	Sud-Est	Mulhouse	16.65	Est
Champigny/Marne	15.09	Ile-de-France	Nancy	18.21	Est
Charleville-Mézières	17.30	Est	Nanterre	6.13	Ile-de-France
Châteauroux	17.37	Centre	Nantes	21.13	Ouest
Cholet	14.00	Ouest	Neuilly-sur-Seine	3.68	Ile-de-France
Clermont-Ferrand	15.85	Centre	Nice	19.75	Sud-Est
Colmar	16.31	Est	Nimes	30.23	Sud-Ouest
Colombes	14.16	Ile-de-France	Niort	19.19	Centre
Courbevoie	4.86	Ile-de-France	Noisy-le-Grand	16.91	Ile-de-France
Créteil	17.58	Ile-de-France	Orléans	20.05	Centre
Dijon	18.75	Centre	Paris	9.15	Ile-de-France
Drancy	10.42	Ile-de-France	Pau	21.31	Sud-Ouest
Dunkerque	28.69	Nord	Perpignan	15.87	Sud-Ouest
Evreux	21.27	Ouest	Pessac	20.71	Sud-Ouest
Fontenay-sous-Bois	12.10	Ile-de-France	Poitiers	21.55	Centre
Grenoble	19.43	Sud-Est	Quimper	16.67	Ouest
Ivry-sur-Seine	9.16	Ile-de-France	Reims	14.98	Est
La Rochelle	18.75	Centre	Rennes	21.75	Ouest
La Seyne-sur-Mer	25.98	Sud-Est	Roubaix	27.97	Nord
Laval	19.48	Ouest	Rouen	20.97	Ouest
Le Havre	17.67	Ouest	Rueil-Malmaison	14.93	Ile-de-France
Le Mans	17.54	Ouest	Saint-Denis	9.17	Ile-de-France
Lille	36.17	Nord	Saint-Etienne	19.90	Sud-Est
Limoges	17.24	Centre	St-Maur-des-Fossés	10.82	Ile-de-France
Lorient	16.74	Ouest	Saint-Nazaire	16.36	Ouest
Lyon	19.09	Sud-Est	Saint-Quentin	20.46	Nord

Ville	Taux taxe d'habitation	Zone Géographique	Ville	Taux taxe d'habitation	Zone Géographique
Sarcelles	19.32	Ile-de-France	Troyes	18.11	Est
Sartrouville	12.38	Ile-de-France	Valence	16.25	Sud-Est
Strasbourg	22.04	Est	Venissieux	18.70	Sud-Est
Toulon	19.37	Sud-Est	Versailles	8.95	Ile-de-France
Toulouse	19.23	Sud-Ouest	Villeneuve-d'Asq	29.96	Nord
Tourcoing	33.61	Nord	Villeurbanne	19.85	Sud-Est
Tours	20.79	Centre	Vitry-sur-Seine	11.50	Ile-de-France

On en déduit pour la variable taux de taxe d'habitation, le tableau suivant obtenu après mise en classes d'amplitudes égales à 5, qui permet déjà de mieux comprendre le phénomène : on voit clairement une concentration des valeurs (84 %) dans l'intervalle [10 ; 25].

TABLEAU 5.3

Classe	Limite infér.	Limite supér.	Point central	Effectif	Fréquence	Effectif cumulé	Fréquence cumulée
1	0.0	5.0	2.5	2	0.0200	2	0.0200
2	5.0	10.0	7.5	6	0.0600	8	0.0800
3	10.0	15.0	12.5	17	0.1700	25	0.2500
4	15.0	20.0	17.5	47	0.4700	72	0.7200
5	20.0	25.0	22.5	20	0.2000	92	0.9200
6	25.0	30.0	27.5	5	0.0500	97	0.9700
7	30.0	35.0	32.5	2	0.0200	99	0.9900
8	35.0	40.0	37.5	1	0.0100	100	1.0000

Dans d'autres cas, on peut recourir à des classes d'amplitudes inégales. ■

5.2 REPRÉSENTATIONS GRAPHIQUES

5.2.1 Barres et camemberts

Pour des variables qualitatives à modalités non ordonnées, il existe une grande variété de diagrammes. Les plus répandus sont :

- les diagrammes en barres (verticales ou horizontales) : les barres sont de longueurs proportionnelles aux fréquences des catégories, leur épaisseur est sans importance.
- Les camemberts (en anglais ***pie-chart***) : chaque catégorie est représentée par une portion de superficie proportionnelle à sa fréquence.



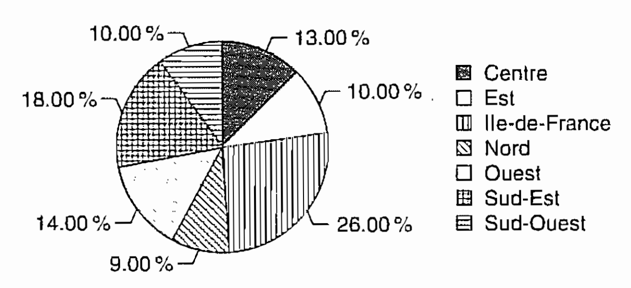


FIGURE 5.1

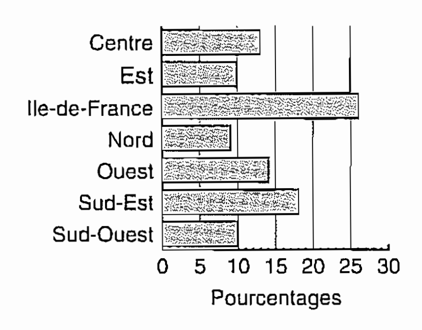


FIGURE 5.2

Pour des variables numériques discrètes, on utilisera de préférence un diagramme en barres verticales comme celui-ci :

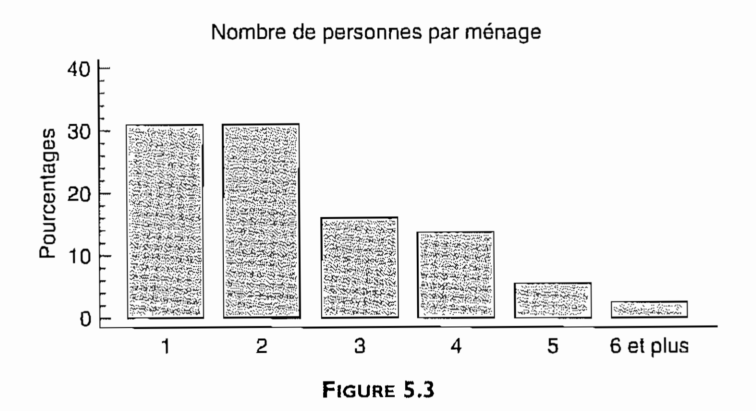


FIGURE 5.3

5.2.2 Histogrammes

Analogues à la courbe de densité d'une variable aléatoire, un histogramme est un graphique à barres verticales accolées, obtenu après découpage en classes des observations d'une variable continue. La surface de chaque barre, encore appelée tuyau d'orgue, doit être proportionnelle à la fréquence de la classe. Pour des classes d'égale amplitude, la hauteur de chaque barre est proportionnelle à la fréquence.

Voici quelques histogrammes de la distribution des taux de taxe d'habitation : tous ont pour propriété que la surface sous l'histogramme vaut 1.

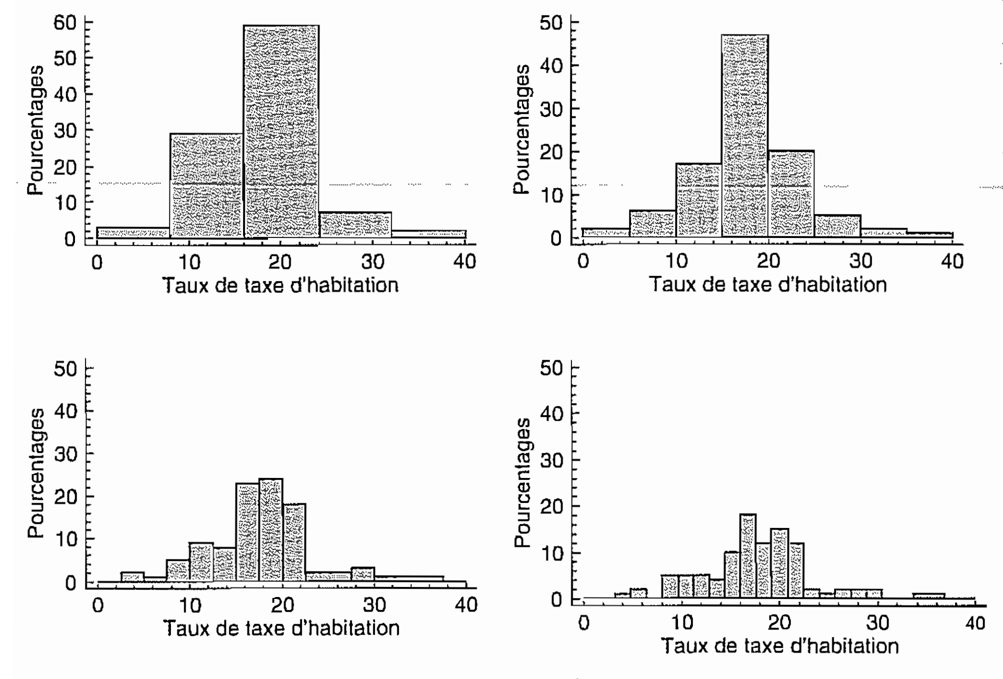


FIGURE 5.3 bis

On constate qu'un trop grand nombre de classes « brouille » l'information.

La détermination du nombre de classes d'un histogramme est délicate et on ne dispose pas de règles absolues. Un trop faible nombre de classes fait perdre de l'information et aboutit à gommer les différences pouvant exister entre des groupes de l'ensemble étudié. En revanche un trop grand nombre de classes aboutit à des graphiques incohérents : certaines classes deviennent vides ou presque, car n est fini.

On peut d'ailleurs critiquer le fait de représenter par une fonction en escalier la distribution d'une variable continue : l'histogramme est une approximation assez pauvre d'une fonction de densité et il serait plus logique de chercher une fonction plus régulière.

La théorie de l'estimation de densité permet de proposer des solutions à ce problème (voir chapitre 13, paragraphe 13.9.3).

Une estimation de densité calculée pour 100 abscisses par la méthode du noyau (ici un noyau cosinus avec une largeur de fenêtre égale à 60 % de l'étendue) fournit une information plus claire, et la forme de la courbe suggère une distribution gaussienne.

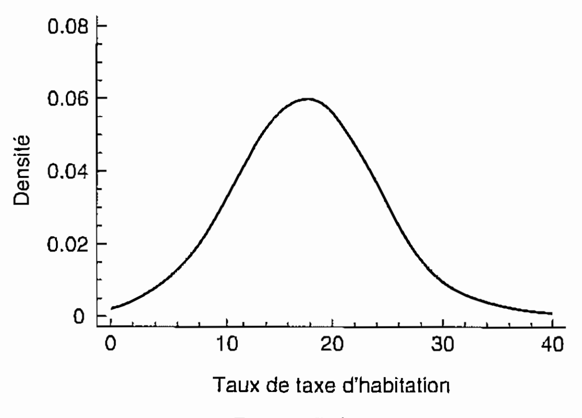


FIGURE 5.4

Mieux qu'un histogramme, une courbe de densité estimée permet des détecter des modes multiples, correspondant à des mélanges de distribution (données provenant de plusieurs populations différentes).

5.2.3 Boîte à moustaches ou box-plot

Ce diagramme, introduit par J.W. Tukey, est une représentation synthétique extrêmement efficace des principales caractéristiques d'une variable numérique. Il en existe plusieurs variantes, mais celle décrite ci-dessous est la plus complète.

La boîte correspond à la partie centrale de la distribution : la moitié des valeurs comprises entre le premier et le troisième quartile Q_1 et Q_3 (voir plus loin). Les moustaches s'étendent de part et d'autre de la boîte jusqu'aux valeurs suivantes : à gauche jusqu'à $Q_1 - 1.5(Q_3 - Q_1)$ si il existe des valeurs encore plus petites, sinon jusqu'à la valeur minimale ; à droite jusqu' à $Q_1 + 1.5(Q_3 - Q_1)$ si il existe des valeurs au-delà, sinon jusqu'à la valeur maximale. Les valeurs au-delà des moustaches repérées par des * sont des valeurs hors norme éventuellement suspectes ou aberrantes mais pas nécessairement.

Ainsi le diagramme en boîte à moustaches montre clairement l'existence de points atypiques pour le taux de taxe d'habitation, ici 3 valeurs très basses, et 4 valeurs très élevées. Il devient alors intéressant d'identifier les individus correspondants.

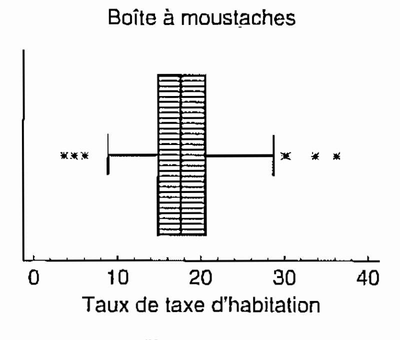


FIGURE 5.5

Un autre grand intérêt de ces diagrammes est de pouvoir faire facilement des comparaisons entre sous-groupes de données : il est plus simple de comparer des diagrammes en boîte que des histogrammes. La figure suivante permet de comparer les distributions du taux de taxe d'habitation selon la région :

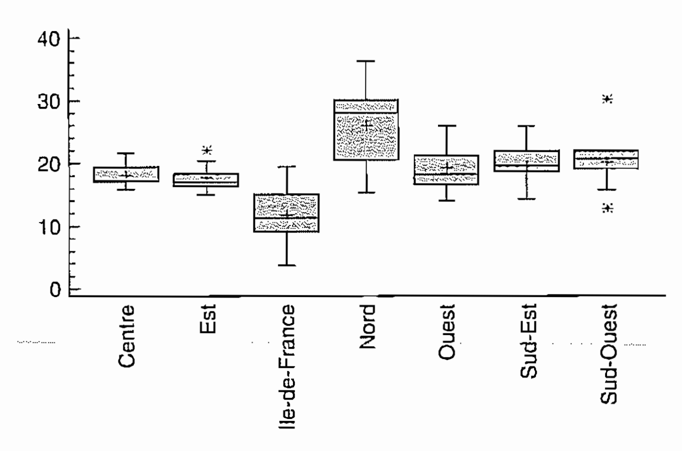


FIGURE 5.6

5.2.4 Courbe de concentration

Applée également courbe de Lorenz, elle est utilisée principalement en statistique économique pour étudier les inégalités de répartition d'une grandeur positive cumulable (revenu, chiffre d'affaire, ...) (fig. 5.7).

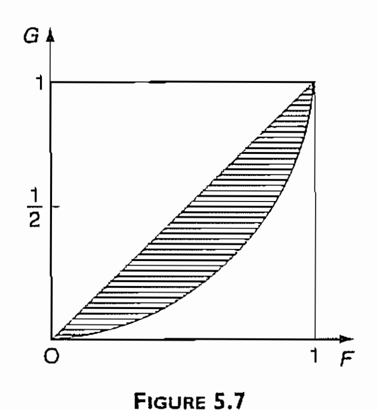


FIGURE 5.7

Exemple : Soit une distribution de revenus X et soit M la masse totale des revenus. A chaque valeur du revenu X, on associe un point de coordonnées $F(x)$ en abscisse : proportion des individus gagnant moins de x et $G(x)$ en ordonnée où $G(x)$ représente la proportion de M correspondante, c'est-à-dire le rapport :

$$\frac{\text{Masse des revenus} < x}{\text{Masse totale}}$$

Cette courbe est toujours en dessous de la première bissectrice car $F(x) > G(x)$ pour une distribution non dégénérée : il suffit de remarquer que les individus qui gagnent moins de x, qui sont donc en proportion $F(x)$, ne peuvent gagner globalement autant que les $100F(x)$ % suivants. ■

La **médiale** $\mathcal{M}$ est la valeur de la variable qui partage en deux la masse totale de la variable. On a donc :

$$\text{Médiale} > \text{Médiane}$$

5.2.4.1 *Propriétés mathématiques*

Supposons connue la distribution théorique de X de densité $f(x)$. L'abscisse d'un point de la courbe est :

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

L'ordonnée correspondante est :

$$q = \frac{\int_{-\infty}^{x} t f(t)\,\mathrm{d}t}{E(X)} = \frac{\int_{-\infty}^{x} t f(t)\,\mathrm{d}t}{\int_{-\infty}^{\infty} t f(t)\,\mathrm{d}t}$$

Si X est une variable qui prend ses valeurs entre x_{min} et x_{max} la courbe de concentration est donc définie en coordonnées paramètriques :

$$F = \int_{x_{\min}}^{x} f(t)\,\mathrm{d}t \qquad \frac{\mathrm{d}F}{\mathrm{d}x} = f(x)$$

$$q = \frac{1}{m}\int_{x_{\min}}^{x} t f(t)\,\mathrm{d}t \qquad \frac{\mathrm{d}q}{\mathrm{d}x} = \frac{1}{m} x f(x)$$

On a : $$\frac{\mathrm{d}q}{\mathrm{d}F} = \frac{\mathrm{d}q}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}F} = \frac{x}{m}$$

On remarque que $\frac{\mathrm{d}q}{\mathrm{d}F} = 1$ si $x = m$.

La courbe possède alors une tangente parallèle à la première bissectrice.

Aux extrémités du carré les pentes des tangentes sont $\frac{x_{\min}}{m}$ et $\frac{x_{\max}}{m}$ respectivement.

Si X varie de 0 à ∞ en particulier, les pentes sont 0 et ∞ (tangente horizontale au départ, verticale à l'arrivée).

5.2.4.2 *Indice de concentration ou indice de Gini*

Plus la distribution de X est inégalement répartie, plus la courbe s'éloigne de la première bissectrice (distribution ultra concentrée : cas où les 9/10 des individus représentent moins de 1/10 de la masse et où le 1/10 restant concentre la quasi-totalité de la variable).

Un indice de concentration proposé par Gini est le double de la surface comprise entre la courbe et la bissectrice (fig. 5.8).

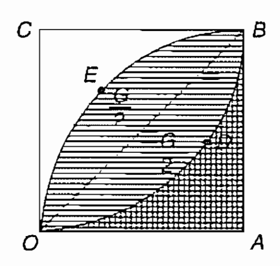


FIGURE 5.8

- Indice de Gini théorique

En prenant la courbe symétrique par rapport à la diagonale, on a :

$$G = \text{aire } OEBA - \text{aire } ODBA \;;$$

$$G = \int_0^1 F\,\mathrm{d}q - \int_0^1 q\,\mathrm{d}F \;;$$

en multipliant par m les deux membres :

$$mG = \int_{-\infty}^{+\infty} F(x)xf(x)\,\mathrm{d}x - m\int_{-\infty}^{+\infty} q(x)f(x)\,\mathrm{d}x$$

$$mG = \int_{-\infty}^{+\infty} x\left[\int_{-\infty}^{x} f(y)\,\mathrm{d}y\right]f(x)\,\mathrm{d}x - \int_{-\infty}^{+\infty}\left[\int_{-\infty}^{x} yf(y)\,\mathrm{d}y\right]f(x)\,\mathrm{d}x$$

$$mG = \int_{-\infty}^{+\infty}\int_{-\infty}^{x}(x-y)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y$$

Comme $\displaystyle\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-y)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y = 0$, il vient :

$$mG = \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{x}(x-y)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y$$

$$+\frac{1}{2}\int_{-\infty}^{+\infty}\int_{x}^{\infty}(y-x)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y$$

Soit : $$mG = \frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|x-y|f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}\Delta_1$$

Δ_1 s'appelle la différence moyenne, d'où :

$$G = \frac{\Delta_1}{2m}$$

- Indice de Gini d'un échantillon

Si toutes les valeurs x_i de la distribution sont distinctes, la différence moyenne empirique vaut :

$$\Delta_1 = \frac{1}{n(n-1)} \sum_i \sum_j |x_i - x_j|$$

$$= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} |x_i - x_j|$$

d'où :

$$\boxed{G = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} |x_i - x_j|}{n(n-1)\bar{x}}}$$

5.3 RÉSUMÉS NUMÉRIQUES

Il est indispensable en général de résumer une série d'observations par des indicateurs typiques dont le plus connu est la moyenne arithmétique. ***Il est cependant toujours insuffisant de résumer une série par un seul indicateur.***

Voici une liste typique de résumés numériques pour la variable taux de taxe d'habitation

TABLEAU 5.4

Effectif	=	100
Moyenne	=	17.7707
Médiane	=	17.625
Variance	=	30.2707
Écart-type	=	5.5019
Minimum	=	3.68
Maximum	=	36.17
Étendue	=	32.49
1er quartile	=	15.035
3e quartile	=	20.585
Intervalle inter-quartiles	=	5.55
Asymétrie	=	0.368299
Aplatissement	=	4.46798
Coef. de variation	=	31.1164 %

5.3.1 Caractéristiques de tendance centrale

Il s'agit en quelque sorte de définir une valeur c autour de laquelle se répartissent les n observations (fig. 5.9).

Les plus usitées sont la médiane, la moyenne arithmétique et le mode.

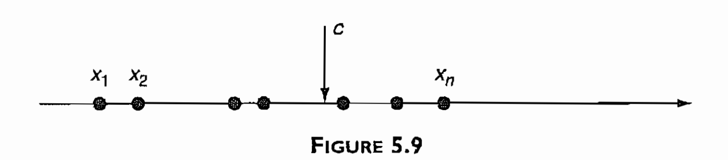


FIGURE 5.9

5.3.1.1 La médiane

C'est la valeur M telle que $F(M) = 0.50$. Si les observations sont rangées par ordre croissant $M = x_{(n+1)/2}$ pour n impair. Si n est pair on prendra conventionnellement :

$$M = \frac{x_{n/2} + x_{n/2+1}}{2}$$

Lorsque l'on ne connaît qu'une répartition en classes (situation à éviter mais que l'on rencontre si l'on travaille sur des documents résultant d'un traitement préalable) on cherche la classe médiane $[e_{i-1}, e_i]$ telle que :

$$F(e_{i-1}) < 0.5 \quad \text{et} \quad F(e_i) > 0.5$$

et on détermine M par interpolation linéaire :

$$M = e_{i-1} + a_i \frac{0.5 - F_{i-1}}{f_i}$$

L'interpolation linéaire revient à supposer une distribution uniforme à l'intérieur de la classe médiane.

La médiane est un indicateur de position insensible aux variations des valeurs extrêmes (elle ne dépend en fait que des valeurs centrales de l'échantillon étudié) mais n'a que peu de propriétés algébriques.

5.3.1.2 La moyenne arithmétique

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{ou} \quad \bar{x} = \sum_{i=1}^{n} p_i x_i \quad \text{pour des données pondérées}$$

Pour des données réparties en k classes la formule $\sum_{i=1}^{k} f_i c_i$ avec $c_i = \frac{e_{i-1} + e_i}{2}$ donne une approximation de la moyenne $\bar{x}$ mais cette situation doit être évitée.

La moyenne arithmétique est fonction de toutes les observations mais est sensible aux valeurs extrêmes : c'est un indicateur peu « robuste » bien qu'universellement utilisé.

5.3.1.3 Le mode

Valeur la plus fréquente pour une distribution discrète ; classe correspondant au pic de l'histogramme pour une variable continue. Sa détermination est malaisée et dépend du découpage en classes.

Pour une répartition parfaitement symétrique on a :

$$\text{Moyenne} = \text{mode} = \text{médiane}$$

5.3.2 Caractéristiques de dispersion

Plus encore que la tendance centrale, la dispersion est la notion clé en statistique car si tous les individus avaient la même valeur il n'y aurait plus de raisonnement statistique. . .

5.3.2.1 L'étendue ou intervalle de variation

$$w = x_{max} - x_{min}$$

Dépendante des valeurs extrêmes c'est un indicateur instable.

5.3.2.2 L'intervalle interquartile

Les quartiles Q_1, Q_2, Q_3 étant définis par $F(Q_1) = 0.25$ $F(Q_2) = 0.50$ et $F(Q_3) = 0.75$, $|Q_3 - Q_1|$ est un indicateur parfois utilisé pour mesurer la dispersion : il est plus robuste que l'étendue.

5.3.2.3 La variance et l'écart-type

Ce sont les deux mesures les plus fréquemment utilisées.

La variance s^2 est définie par :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{ou} \qquad \sum p_i(x_i - \bar{x})^2$$

L'écart-type s s'exprime dans la même unité que la variable étudiée.

Le coefficient de variation exprime en pourcentage le rapport $\frac{s}{\bar{x}}$. Il n'a de sens que si $\bar{x} > 0$.

On a les formules suivantes :

$$s^2 = \frac{1}{n}\sum x_i^2 - (\bar{x})^2 = \text{(moyenne des carrés) moins (carré de la moyenne) ;}$$

$$s^2 = \frac{1}{n}\sum(x_i - a)^2 - (\bar{x} - a)^2, \text{ théorème de König-Huyghens.}$$

Ces deux formules ne présentent d'intérêt que pour des calculs à la main sur des petites séries et doivent être prohibées pour des calculs automatiques sur des grandes séries, les sommes de carrés pouvant conduire à des dépassements de capacité ou à des pertes de précision.

L'algorithme suivant permet de calculer la somme des carrés des écarts SC à la moyenne pour n valeurs par ajustement progressif : chaque nouvelle valeur x_j introduite entraîne une modification simple et positive de la somme des carrés calculée pour les $j - 1$ valeurs déjà introduites :

$$SC = 0$$

$$T = x_1$$

pour $j = 2, 3, \ldots, n$ faire :

$$T = T + x_j$$

$$SC = SC + \frac{1}{j(j-1)}(jx_j - T)^2$$

d'où quand $j = n$, $\bar{x} = T/n$ et $s^2 = SC/n$.

5.3.3 Cohérence entre tendance centrale et dispersion

Nous pouvons considérer qu'une valeur centrale c doit être « proche » de l'ensemble des x_i et minimiser une fonction du type $\frac{1}{n}\sum_{i=1}^{n} d(c\,;x_i)$ où d est un écart. $D = \frac{1}{n}\sum d(c\,;x_i)$ définit alors une mesure de dispersion des observations autour de c.

Le choix d'une certaine forme analytique pour d entraîne alors l'usage simultané d'une mesure de tendance centrale et d'une mesure de dispersion cohérentes :

- si $d(c\,;x_i) = (c - x_i)^2$ on a $c = \bar{x}$ et $D = s^2$;
- si $d(c\,;x_i) = |c - x_i|$ on trouve $c = M$, c'est-à-dire la médiane et $D = \frac{1}{n}\sum|x_i - M|$.

Le couple $(\bar{x}, s^2)$ est donc cohérent avec l'usage de distances euclidiennes.

Géométriquement si l'ensemble des observations de X est considéré comme un vecteur de $\mathbb{R}^n$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

chercher une valeur centrale c revient à chercher une variable constante c'est-à-dire un vecteur :

$$\mathbf{c} = c\begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} = c\mathbf{1}$$

le plus proche possible de **x** au sens d'une certaine topologie.

En munissant $\mathbb{R}^n$ de la métrique euclidienne usuelle, $\bar{x}$ est la mesure de la projection de **x** sur Δ (fig. 5.10).

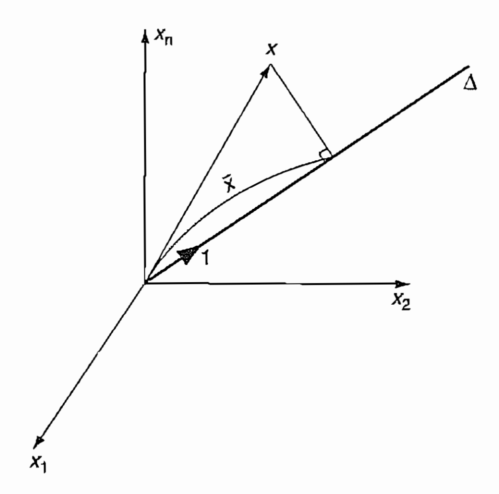


FIGURE 5.10

5.3.4 Caractéristiques de forme

Elles sont utiles notamment pour vérifier rapidement le caractère normal d'une distribution (on sait que le coefficient d'aplatissement théorique de la loi normale est 3) :

- coefficient d'asymétrie : $\gamma_1 = m_3/s^3$;
- coefficient d'aplatissement : $\gamma_2 = m_4/s^4$;

où $m_3 = \dfrac{1}{n}\sum(x_i - \bar{x})^3$ et $m_4 = \dfrac{1}{n}\sum(x_i - \bar{x})^4$.

Les notations ne sont pas universelles et γ_1 est parfois noté $\sqrt{b_1}$, γ_2 noté b_2. Certains auteurs utilisent $\gamma_2 - 3$.

6 Description bidimensionnelle et mesures de liaison entre variables

Après les descriptions unidimensionnelles on étudie généralement les liaisons entre les variables observées : c'est ce que l'on appelle communément l'étude des corrélations. Les méthodes et les indices de dépendance varient selon la nature (qualitative, ordinale, numérique) des variables étudiées.

6.1 LIAISON ENTRE DEUX VARIABLES NUMÉRIQUES

Supposons que l'on observe pour n individus deux variables X et Y. On a donc n couples $(x_i ; y_i)$ ou encore deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^n$ avec :

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

6.1.1 Étude graphique de la corrélation

Afin d'examiner s'il existe une liaison entre X et Y on représente chaque observation i comme un point de coordonnées (x_i, y_i) dans un repère cartésien. La forme du nuage de points ainsi tracé est fondamentale pour la suite : ainsi la figure 6.1 montre :

a) une absence de liaison ;
b) une absence de liaison en moyenne mais pas en dispersion ;
c) une corrélation linéaire positive ;
d) une corrélation non linéaire.

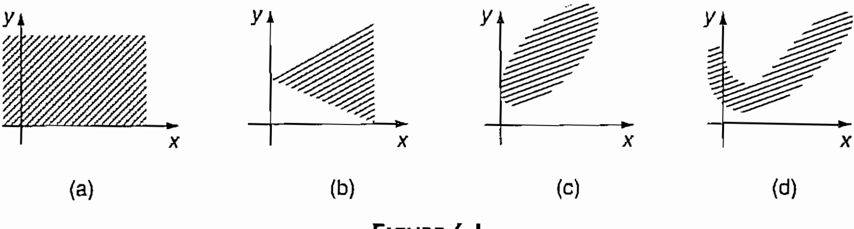


FIGURE 6.1

On dit qu'il y a corrélation si il y a dépendance en moyenne : à $X = x$ fixé la moyenne $\overline{Y}$ est fonction de x. Si cette liaison est approximativement linéaire on se trouve dans le cas de la corrélation linéaire.

Rappelons que la ***non corrélation n'est pas nécessairement l'indépendance.***

6.1.2 Le coefficient de corrélation linéaire

Ce coefficient dit de « Bravais-Pearson » mesure exclusivement le caractère plus ou moins linéaire du nuage de points.

6.1.2.1 *Définition*

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

où s_x et s_y sont les écarts-types de x et y :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Le numérateur $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ est la covariance observée.

De même que pour ρ (voir chapitre 3) dont il est la version empirique : $-1 \leq r \leq 1$ et $|r| = 1$ est équivalent à l'existence d'une relation linéaire exacte : $ax_i + by_i + c = 0 \quad \forall i$.

Si l'on considère dans l'espace $\mathbb{R}^n$ les deux vecteurs :

$$\begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \cdot \\ \cdot \\ \cdot \\ x_n - \bar{x} \end{bmatrix} \quad \text{et} \quad \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \cdot \\ \cdot \\ \cdot \\ y_n - \bar{y} \end{bmatrix}$$

r est le cosinus de l'angle formé par ces deux vecteurs comme un calcul élémentaire le montre, d'où ses propriétés.

6.1.2.2 Du bon usage du coefficient r

r ne mesure que le caractère ***linéaire*** d'une liaison et son usage doit être réservé à des nuages où les points sont répartis de part et d'autre d'une tendance linéaire (fig. 6.1c du paragraphe précédent).

Par contre, la figure 6.2[1] montre les risques d'un usage inconsidéré du coefficient de corrélation linéaire r. On notera en particulier que r est très sensible aux individus extrêmes et n'est donc pas « robuste ».

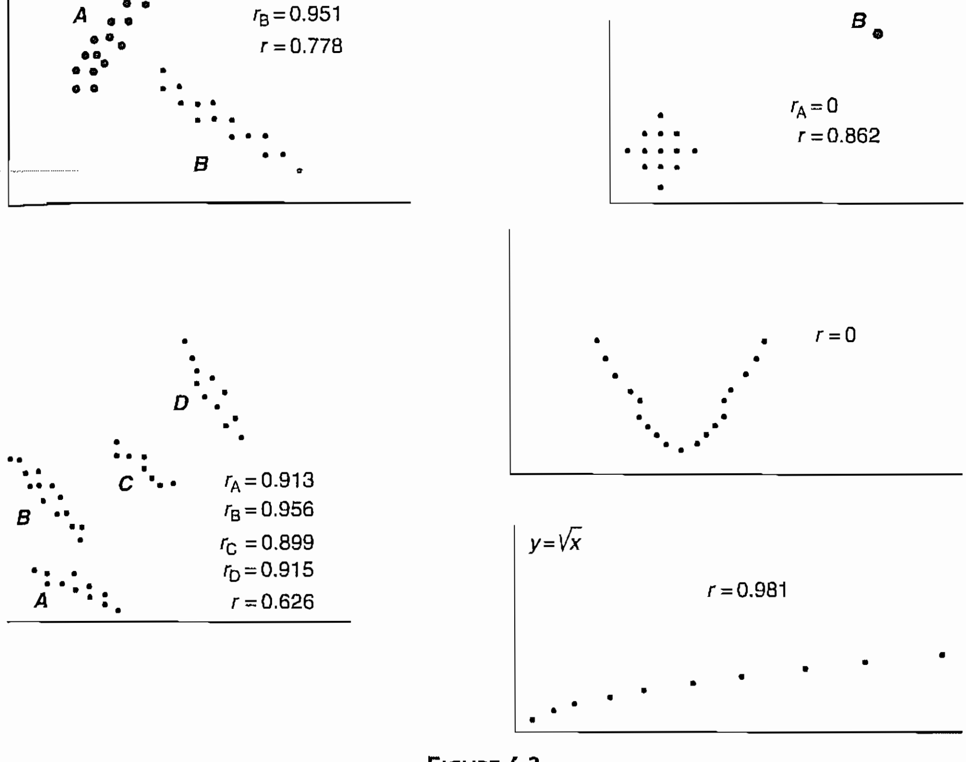


FIGURE 6.2

Les 4 nuages de la figure 6.3 ont mêmes moyennes, mêmes variances et même coefficient de corrélation :

$$\bar{x} = 9 \qquad \bar{y} = 7.5$$
$$s_x^2 = 10.0 \qquad s_y^2 = 3.75$$
$$r = 0.82$$

Seul le premier nuage justifie l'usage de r.

1 Cailliez et Pagès, 1976.

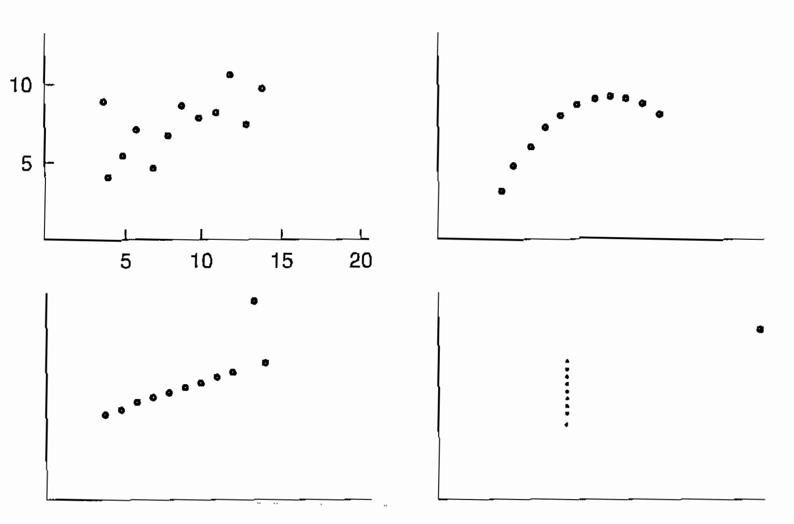


Figure 6.3

Notons pour finir que la corrélation n'est pas transitive : **x** très corrélé avec **y**, **y** très corrélé avec **z**, n'implique nullement que **x** soit corrélé avec **z**.

6.1.2.3 Matrice de corrélation entre p variables

Lorsque l'on observe les valeurs numériques de p variables sur n individus on se trouve en présence d'un tableau **X** à n lignes et p colonnes :

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 2 \\ . \\ . \\ i \\ . \\ . \\ . \\ n \end{array} \begin{array}{c} 1 \quad 2 \quad \dots \quad j \quad \dots \quad p \\ \left[\begin{array}{ccccccc} & & & . & & & \\ & & & . & & & \\ & & & . & & & \\ & & & . & & & \\ . & . & \dots & x_i^j & \dots & . & \\ & & & . & & & \\ & & & . & & & \\ & & & . & & & \\ & & & . & & & \end{array}\right] \end{array}$$

x_i^j est la valeur prise par la variable n° j sur le $i^{\text{ème}}$ individu.

Le tableau des données centrées **Y** s'obtient en utilisant l'opérateur de centrage $\mathbf{A} = \mathbf{I} - \dfrac{\mathbf{11}'}{n}$.

$\mathbf{Y} = \mathbf{AX}$.

$\mathbf{A}$ est la matrice $n \times n$ de terme général :

$$a_{ii} = 1 - \frac{1}{n}, \qquad a_{ij} = -\frac{1}{n} \quad \text{si} \quad i \neq j.$$

La matrice des variances et covariances des p variables :

$$\mathbf{V} = \begin{bmatrix} s_1^2 & s_{12} & \cdot\cdot & s_{1p} \\ \cdot & s_2^2 & \cdot\cdot & s_{2p} \\ & & & \cdot \\ & & & \cdot \\ \cdot & & & s_p^2 \end{bmatrix}$$

où :

$$s_{kl} = \frac{1}{n} \sum_{i=1}^{n} x_i^k x_i^l - \bar{x}^k \bar{x}^l$$

est telle que $\mathbf{V} = \dfrac{1}{n}\mathbf{Y}'\mathbf{Y}$.

La matrice regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux est notée $\mathbf{R}$:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \cdot & 1 & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ r_{p1} & & & 1 \end{bmatrix}$$

En posant :

$$\mathbf{D}_{1/s} = \begin{bmatrix} 1/s_1 & & \cdots & 0 \\ \cdot & 1/s_2 & \ddots & \vdots \\ \cdot & & & \\ 0 & & & 1/s_p \end{bmatrix}$$

On a $\mathbf{R} = \mathbf{D}_{1/s}\mathbf{V}\mathbf{D}_{1/s}$.

$\mathbf{R}$ est identique à la matrice de variance-covariance des données centrées et réduites.

$\mathbf{R}$ résume la structure des dépendances linéaires entre les p variables.

Comme $\mathbf{V}$, $\mathbf{R}$ est une matrice symétrique positive.

Exemple : Les résultats suivants concernent 6 variables du tableau de données figurant au chapitre 17 et donnant pour 18 véhicules des caractéristiques techniques.

La matrice **V** est calculée avec $n - 1$ en dénominateur :

Matrice de variance et covariance **V**

	CYL	PUIS	LON	LAR	POIDS	VITESSE
CYL	139823.5294	6069.7451	5798.7059	1251.2941	40404.2941	3018.5686
PUIS	6069.7451	415.1928	288.9118	56.3922	2135.6961	208.8791
LON	5798.7059	288.9118	488.7353	99.7647	2628.3824	127.7353
LAR	1251.2941	56.3922	99.7647	28.2353	521.7059	30.5098
POIDS	40404.2941	2135.6961	2628.3824	521.7059	18757.4412	794.1078
VITESSE	3018.5686	208.8791	127.7353	30.5098	794.1078	147.3889

La matrice **R** est la suivante :

Matrice de corrélation **R** (Bravais-Pearson)

	CYL	PUIS	LON	LAR	POIDS	VITESSE
CYL	1.00000	0.79663	0.70146	0.62976	0.78895	0.66493
PUIS	0.79663	1.00000	0.64136	0.52083	0.76529	0.84438
LON	0.70146	0.64136	1.00000	0.84927	0.86809	0.47593
LAR	0.62976	0.52083	0.84927	1.00000	0.71687	0.47295
POIDS	0.78895	0.76529	0.86809	0.71687	1.00000	0.47760
VITESSE	0.66493	0.84438	0.47593	0.47295	0.47760	1.00000

On constate que toutes les variables sont corrélées positivement, avec certains coefficients très élevés : il existe donc une forte redondance entre les variables, ce qui conduit à un phénomène dit de multicolinéarité (voir chapitre 17).

La figure suivante, appelée matrice de dispersion, est très utile : elle permet en un seul graphique de juger des liaisons entre toutes les variables.

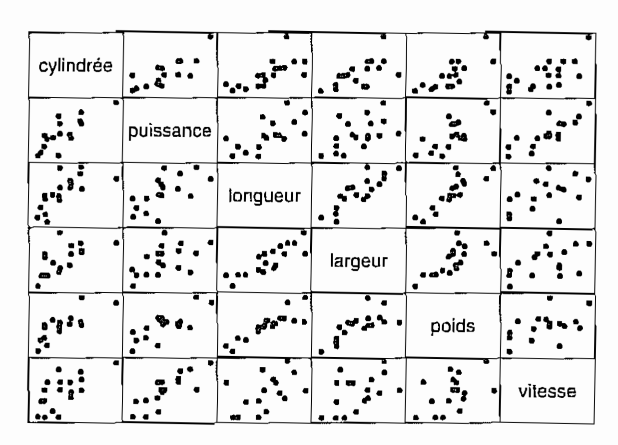


FIGURE 6.4 Matrice de dispersion

6.1.3 Caractère significatif d'un coefficient de corrélation

En admettant que l'on se trouve dans le cas où l'usage de r est justifié, à partir de quelle valeur la liaison est-elle significative ?

En anticipant sur la théorie des tests on raisonne comme suit : si les n observations avaient été prélevées au hasard dans une population où X et Y sont indépendantes (donc où $\rho = 0$) quelle seraient les valeurs possibles de r ou plus exactement la distribution de probabilité de la variable R qui correspond à cet échantillonnage ?

Lorsque $\rho = 0$ et que les observations proviennent d'un couple gaussien la distribution de R est relativement facile à obtenir.

On montre que : $$\boxed{\frac{R}{\sqrt{1 - R^2}}\sqrt{n - 2} \text{ suit une loi } T_{n-2}}$$

Par changement de variable on en déduit alors directement la densité de R si $\rho = 0$:

$$f(r) = \frac{1}{B\left(\frac{1}{2} ; \frac{n-2}{2}\right)}(1 - r^2)^{(n-4)/2}$$

Pour $n = 4$, on remarquera que R suit une loi uniforme sur $[-1, 1]$ et donc que toutes les valeurs possibles sont équiprobables.

On a : $$\boxed{E(R) = 0 \quad \text{et} \quad V(R) = \frac{1}{n - 1}}$$

Pour $n > 100$, la loi de R est approximée de très près par une loi de Laplace-Gauss :

$$LG\left(0 ; \frac{1}{\sqrt{n - 1}}\right)$$

Sinon la loi de R est tabulée, Table A.9.

Ainsi au risque 5 % on déclarera qu'une liaison est significative sur un échantillon de 30 observations si $|r| > 0.36$.

On remarquera que le seuil de signification décroît quand n croît ; le fait de trouver que r diffère significativement de 0 ne garantit nullement que la liaison soit forte (voir chapitre 16).

Lorsque ρ est différent de zéro la loi exacte de R bien que connue est très difficilement exploitable on notera cependant que :

$$E(R) \simeq \rho - \frac{\rho(1 - \rho^2)}{2n} \qquad R \text{ est biaisé pour } \rho$$

$$V(R) \simeq \frac{(1 - \rho^2)^2}{n - 1}$$

La figure 6.5 donne les distributions d'échantillonnage de r pour différentes valeurs de ρ, avec $n = 10$. On ne peut pas faire directement une approximation normale. On utilisera plutôt le résultat suivant conduisant à une approximation correcte dès que $n > 25$.

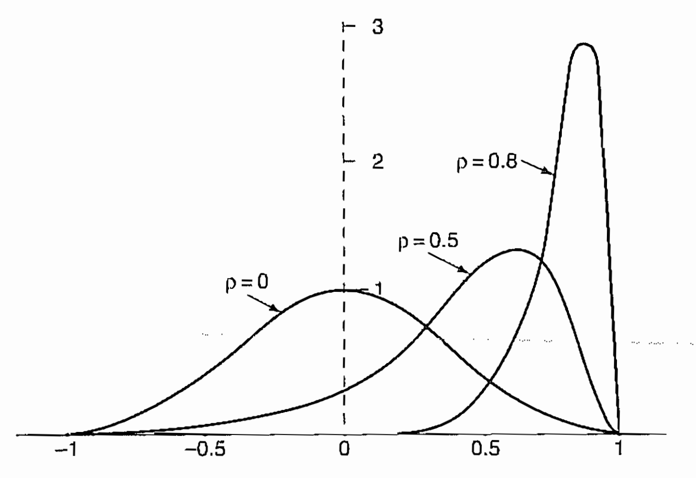


FIGURE 6.5

$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \xrightarrow[n\to\infty]{\mathcal{L}} LG\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}\,;\,\frac{1}{\sqrt{n-3}}\right)$$

Z est la transformée de Fisher de R (table A.10). On notera que $V(Z)$ est indépendant de ρ. Cette transformation permet de tester des valeurs *a priori* pour ρ et de trouver des intervalles de confiance pour ρ à partir de R. On peut également utiliser l'abaque fourni en annexe (table A1.9 *bis*).

Lorsque le couple (X, Y) n'est pas gaussien les résultats précédents restent utilisables ***à condition que n soit grand*** (en pratique $n > 30$), mais le fait de trouver que r n'est pas significativement différent de 0 n'entraîne pas nécessairement l'indépendance.

6.1.4 Corrélation partielle

Il arrive fréquemment que la dépendance apparente entre deux variables soit due en réalité aux variations d'une troisième variable. La littérature statistique abonde en exemple de fausses corrélations surprenantes entre phénomènes variés qui disparaissent lorsque l'on fixe une troisième variable (souvent non aléatoire comme le temps) ainsi de la corrélation entre le nombre de maladies mentales déclarées chaque année et le nombre de postes de radio installés.

Les coefficients de corrélation partielle constituent un moyen d'éliminer l'influence d'une ou plusieurs variables.

Ces coefficients peuvent être introduits de diverses façons en particulier dans le cadre de la régression multiple (chapitre 17). Nous en donnerons ici deux présentations, l'une issue du modèle gaussien, l'autre géométrique.

6.1.4.1 Le modèle normal à p dimensions

Soit un vecteur aléatoire $(X_1, X_2, \ldots X_p)$ suivant une loi $N_p(\boldsymbol{\mu}, \Sigma)$. En appliquant les résultats du chapitre 4, paragraphe 4.2.4, on sait que la loi du couple $X_1 X_2$ conditionnée par $X_3 X_4 \ldots X_p$ est une loi normale à deux dimensions. On obtient alors le coefficient de corrélation partiel (ou conditionnel) $\rho_{12.34\ldots p}$, à partir de la matrice des covariances partielles.

Un calcul simple montre qu'en particulier pour $p = 3$:

$$\rho_{x_1 x_2 . x_3} = \frac{\rho_{x_1 x_2} - \rho_{x_1 x_3}\rho_{x_2 x_3}}{\sqrt{(1 - \rho^2_{x_1 x_3})(1 - \rho^2_{x_2 x_3})}}$$

Cette formule se généralise et permet de calculer de proche en proche les divers coefficients de corrélation partielle :

$$\rho_{x_1 x_2 . x_3 x_4} \qquad \rho_{x_1 x_2 . x_3 x_4 x_5} \cdots$$

Pour obtenir $\rho_{x_1 x_2 . x_3 x_4}$ il suffit de remplacer dans la formule précédente les corrélations simples par les corrélations partielles :

$$\rho_{x_1 x_2 . x_3 x_4} = \frac{\rho_{x_1 x_2 . x_3} - \rho_{x_1 x_4 . x_3}\,\rho_{x_2 x_4 . x_3}}{\sqrt{(1 - \rho^2_{x_1 x_4 . x_3})(1 - \rho^2_{x_2 x_4 . x_3})}}$$

On définit alors formellement le coefficient de corrélation linéaire partielle empirique en remplaçant ρ par r.

6.1.4.2 Corrélation entre résidus

Ayant défini $r_{x_1 x_2 . x_3}$ par les formules précédentes, il est facile de vérifier que ce coefficient mesure la corrélation entre le résidu de l'ajustement linéaire de $\mathbf{x}_1$ sur $\mathbf{x}_3$ et celui de l'ajustement de $\mathbf{x}_2$ sur $\mathbf{x}_3$.

Interprétation géométrique dans $\mathbb{R}^n$ **:** Nous supposerons $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ centrées.

$\hat{\mathbf{y}}_3$ est la projection de $\mathbf{x}_1$ sur $\mathbf{x}_3$.

$\hat{\mathbf{x}}$ est la projection de $\mathbf{x}_1$ sur le plan $\mathbf{x}_2$, $\mathbf{x}_3$ (fig. 6.6).

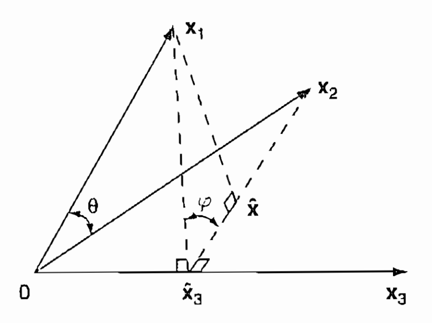


FIGURE 6.6

On a alors $\cos\theta = r_{x_1 x_2}$ et $\cos\varphi = r_{x_1 x_2 . x_3}$ φ est la projection de l'angle entre $\mathbf{x}_1$ et $\mathbf{x}_2$ sur un plan perpendiculaire à x_3.

On peut vérifier ainsi que $r_{x_1x_2.x_3}$ est le coefficient de corrélation linéaire entre la partie de $\mathbf{x}_1$ non expliquée linéairement par $\mathbf{x}_3$ et la partie de $\mathbf{x}_2$ non expliquée linéairement par $\mathbf{x}_3$. On voit que si $\mathbf{x}_3$ est très voisin de $\mathbf{x}_2$ la corrélation partielle est voisine de 0 car $\mathbf{x}_2$ n'apporte presque pas d'information supplémentaire sur $\mathbf{x}_1$ une fois $\mathbf{x}_3$ connu.

6.1.4.3 Signification d'un coefficient de corrélation partielle

Dans le cas gaussien, on démontre que la loi du coefficient de corrélation partielle est la même que celle d'un coefficient de corrélation simple mais avec un degré de liberté diminué de d, nombre de variables fixées.

Donc $\dfrac{r}{\sqrt{1-r^2}}\sqrt{n-d-2}$ suit un T_{n-d-2}, ce qui permet de tester le caractère significatif d'une liaison partielle.

Exemple : (voir les données complètes au chapitre 17 « Régression multiple ») Sur l'échantillon de 18 automobiles, la matrice de corrélation entre prix, vitesse et puissance est :

	Prix	Vitesse	Puissance
Prix	1	0.58176	0.79870
Vitesse	0.58176	1	0.84438
Puissance	0.79870	0.84438	1

Au seuil 5 % toutes ces corrélations sont significatives (valeur critique 0.468).

Cependant, le coefficient de corrélation entre le prix et la vitesse sachant la puissance vaut :

$$\frac{0.58176 - 0.79870 \times 0.84438}{\sqrt{(1-(0.79870)^2)(1-(0.84438)^2)}} = -0.28739$$

La liaison a changé de signe mais elle n'est plus significative (valeur critique à 5 % : 0.482).

6.2 CORRÉLATION MULTIPLE ENTRE UNE VARIABLE NUMÉRIQUE ET *p* AUTRES VARIABLES NUMÉRIQUES

6.2.1 Définition

Soit une variable numérique $\mathbf{y}$ et un ensemble de p variables également numériques $\mathbf{x}^1, \mathbf{x}^2, \ldots \mathbf{x}^p$.

Le coefficient de corrélation multiple R est alors la valeur maximale prise par le coefficient de corrélation linéaire entre $\mathbf{y}$ et une combinaison linéaire des $\mathbf{x}^j$:

$$\boxed{R = \sup_{a_1, a_2, \ldots, a_p} r\left(\mathbf{y}\,;\, \sum_{j=1}^{p} a_j \mathbf{x}^j\right)}$$

On a donc toujours $0 \leq R \leq 1$.

$R = 1$ si il existe une combinaison linéaire des $\mathbf{x}^j$ telle que :

$$\mathbf{y} = a_0 + \sum_{j=1}^{p} a_j \mathbf{x}^j$$

6.2.2 Interprétation géométrique

Rappelons que le coefficient de corrélation est le cosinus de l'angle formé dans $\mathbb{R}^n$ par des variables centrées. R est donc le cosinus du plus petit angle formé par $\mathbf{y}$ (centrée) et une combinaison linéaire des $\mathbf{x}^i$ centrées.

Considérons le sous-espace W de $\mathbb{R}^n$ (de dimension au plus égale à $p + 1$) engendré par les combinaisons linéaires des $\mathbf{x}^j$ et la constante $\mathbf{1}$.

R est alors le cosinus de l'angle θ formé par la variable centrée $\mathbf{y} - \bar{\mathbf{y}}$ et W, c'est-à-dire l'angle formé par $\mathbf{y} - \bar{\mathbf{y}}$ et sa projection orthogonale $\mathbf{y}^* - \bar{\mathbf{y}}$ sur W (fig. 6.7).

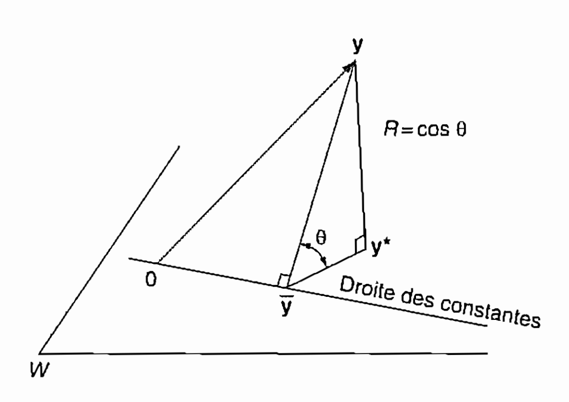


FIGURE 6.7

6.2.3 Calcul de R

Soit $\mathbf{A}$ la matrice de projection orthogonale sur W, alors :

$$R^2 = \frac{(\mathbf{y} - \bar{\mathbf{y}})' \mathbf{A}(\mathbf{y} - \bar{\mathbf{y}})}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} = \frac{s^2_{y^*}}{s^2_y}$$

En particulier si $\mathbf{y}$ est centré :

$$\boxed{R^2 = \frac{\mathbf{y}'\mathbf{A}\mathbf{y}}{\mathbf{y}'\mathbf{y}}}$$

En effet $\|\mathbf{A}\mathbf{y}\|^2 = \cos^2\theta \|\mathbf{y}\|^2$ et $\|\mathbf{A}\mathbf{y}\|^2 = \mathbf{y}'\mathbf{A}'\mathbf{A}\mathbf{y} = \mathbf{y}'\mathbf{A}\mathbf{y}$ car $\mathbf{A}$ est un projecteur orthogonal ($\mathbf{A} = \mathbf{A}'$ et $\mathbf{A}^2 = \mathbf{A}$).

Si **X** désigne la matrice dont les colonnes sont les p variables $\mathbf{x}^1, \mathbf{x}^2 \ldots, \mathbf{x}^p$ centrées et si **y** est centrée :

$$\boxed{R^2 = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}}$$

où $(\mathbf{X}'\mathbf{X})^{-}$ est une inverse généralisée quelconque de $(\mathbf{X}'\mathbf{X})$.

On peut démontrer alors la formule reliant corrélation multiple et corrélations partielles des divers ordres :

$$\boxed{1 - R^2_{y.x_1\ldots x_p} = (1 - r^2_{yx_1})(1 - r^2_{yx_2.x_1})(1 - r^2_{yx_3.x_1x_2})\ldots(1 - r^2_{yx_p.x_1x_2\ldots x_{p-1}})}$$

6.2.4 Signification d'un coefficient de corrélation multiple

Si les n observations étaient issues d'une population gaussienne où Y est indépendante des X^j alors on démontre que (voir chapitre 17) :

$$\boxed{\frac{R^2}{1 - R^2}\frac{n - p - 1}{p} = F(p, n - p - 1)}$$

On retrouve comme cas particulier la loi du coefficient de corrélation linéaire simple en faisant $p = 1$.

6.3 LIAISON ENTRE VARIABLES ORDINALES : LA CORRÉLATION DES RANGS

Il arrive souvent de ne disposer que d'un ordre sur un ensemble d'individus et non de valeurs numériques d'une variable mesurable : soit parce qu'on ne dispose que de données du type classement (ordre de préférence, classement A, B, C, D, E), ou bien parce que les valeurs numériques d'une variable n'ont que peu de sens et n'importent que par leur ordre (notes d'une copie de français : avoir 12 ne signifie pas valoir deux fois plus que celui qui a 6).

A chaque individu de 1 à n on associe son rang selon une variable (un rang varie de 1 à n). Étudier la liaison entre deux variables revient donc à comparer les classements issus de ces deux variables :

Objet :	1	2	....	n
Rang n° 1 :	r_1	r_2		r_n
Rang n° 2 :	s_1	s_2		s_n

Les r_i et s_i sont des permutations différentes des n premiers entiers.

6.3.1 Le coefficient de Spearman

Le psychologue Charles Spearman a proposé en 1904 de calculer le coefficient de corrélation sur les rangs :

$$r_s = \frac{\text{cov}(r, s)}{s_r s_s}$$

Le fait que les rangs soient des permutations de $[1 \ldots n]$ simplifie les calculs et l'on a en l'absence d'ex aequo :

$$\bar{r} = \bar{s} = \frac{n+1}{2} \qquad s^2(r) = s^2(s) = \frac{n^2-1}{12}$$

d'où :

$$r_s = \frac{\frac{1}{n}\sum_i r_i s_i - \left(\frac{n+1}{2}\right)^2}{\frac{n^2-1}{12}}$$

Si l'on pose $d_i = r_i - s_i$ différence des rangs d'un même objet selon les deux classements, on a :

$$\sum_i r_i s_i = -\frac{1}{2}\sum_i (r_i - s_i)^2 + \frac{1}{2}\sum_i r_i^2 + \frac{1}{2}\sum_i s_i^2$$

mais :

$$\sum_i r_i^2 = \sum_i s_i^2 = \frac{n(n+1)(2n+1)}{6}$$

somme des carrés des nombres entiers, d'où :

$$r_s = -\frac{6\sum_i d_i^2}{n(n^2-1)} + \frac{\frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2}{\frac{n^2-1}{12}}$$

Le deuxième terme vaut 1 après calcul et on a la formule pratique :

$$\boxed{r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}}$$

La définition de r_s comme coefficient de corrélation linéaire sur des rangs nous indique que :

$r_s = 1 \Rightarrow$ les deux classements sont identiques ;

$r_s = -1 \Rightarrow$ les deux classements sont inverses l'un de l'autre ;

$r_s = 0 \Rightarrow$ les deux classements sont indépendants.

Pour savoir si la valeur trouvée de r_s est significative, on se reportera à la table du coefficient de corrélation de Spearman fournie en annexe[2]

La région critique sera $|R_s| > k$:

- si $R_s > k$: il y a concordance des classements ;
- si $R_s < -k$: il y a discordance des classements.

Lorsque les observations proviennent d'un couple normal (X, Y) de corrélation ρ et que l'on calcule r_s à la place de r on montre que si n est très grand on a les relations approchées suivantes :

$$r_s \simeq \frac{6}{\pi} \text{Arc sin}\left(\frac{\rho}{2}\right) \qquad \text{ou} \qquad \rho \simeq 2 \sin\left(\frac{\pi}{6} r_s\right)$$

6.3.2 Le coefficient de corrélation des rangs τ de M. G. Kendall

6.3.2.1 *Aspect théorique*

Afin de savoir si deux variables aléatoires X et Y varient dans le même sens ou en sens contraire on peut considérer le signe du produit $(X_1 - X_2)(Y_1 - Y_2)$ où (X_1, Y_1) (X_2, Y_2) sont deux réalisations indépendantes du couple (X, Y).

Si $P((X_1 - X_2)(Y_1 - Y_2) > 0) > 1/2$ il y a plus de chances d'observer une variation dans le même sens que dans le sens inverse.

On définit alors le coefficient théorique τ par :

$$\tau = 2P((X_1 - X_2)(Y_1 - Y_2) > 0) - 1$$

Ce coefficient est donc compris entre -1 et $+1$ et s'annule lorsque X et Y sont indépendantes (mais pas seulement dans ce cas . . .).

Si (X, Y) est un couple gaussien de coefficient de corrélation ρ on montre que :

$$\tau = \frac{2}{\pi} \text{Arc sin}\, \rho$$

On remarquera que $\tau \leq \rho$. $\tau = \rho$ n'est vrai que pour $\rho = 0$ et $\rho = \pm 1$.

Notons enfin que :

$$\tau = P((X_i - X_j)(Y_i - Y_j) > 0) - P((X_i - X_j)(Y_i - Y_j) < 0) = p_c - p_d$$

où p_c et p_d sont respectivement les probabilités de concordance et de discordance.

6.3.2.2 *Calcul sur un échantillon*

En s'inspirant des considérations précédentes :

On considère tous les couples d'individus. On note 1 si deux individus i et j sont dans le même ordre pour les deux variables : $x_i < x_j$ et $y_i < y_j$.

---

2 Cette table est obtenue en utilisant le fait que dans le cas d'indépendance, les $n!$ permutations d'un classement sont équiprobables.

On note -1 si les deux classements discordent $x_i < x_j$ et $y_i > y_j$.

On somme les valeurs obtenues pour les $\frac{n(n-1)}{2}$ couples distincts, soit S cette somme ; on a :

$$S_{\max} = -S_{\min} = \frac{n(n-1)}{2}$$

Le coefficient τ est alors défini par :

$$\boxed{\tau = \frac{2S}{n(n-1)}}$$

On constate que :

$\tau = 1$ classements identiques ;
$\tau = -1$ classements inversés.

Pour savoir si la valeur constatée est significative on se réfère à la situation théorique d'indépendance dans la population.

On peut alors obtenir la distribution de τ par des arguments combinatoires mais celle-ci peut être approchée par une loi de Laplace-Gauss :

$$\tau \sim LG\left(0 ; \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$$

L'approximation est très bonne dès que $n \geq 8$, ce qui est un avantage pratique sur le coefficient de Spearman, si l'on ne dispose pas de tables de ce dernier.

Méthode de calcul rapide : on ordonne les x_i de 1 à n ; on compte pour chaque x_i le nombre de $y_j > y_i$ parmi ceux pour lesquels $j > i$. On somme ; soit R cette somme :

$$S = 2R - \frac{n(n-1)}{2}$$

$$\tau = \frac{4R}{n(n-1)} - 1$$

Exemple : 10 échantillons de cidre ont été classés par ordre de préférence par deux gastronomes :

x_i	1	2	3	4	5	6	7	8	9	10
y_i	3	1	4	2	6	5	9	8	10	7

Le coefficient de Spearman : $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ vaut $r_s = 0.84$

Le coefficient de Kendall s'obtient par :

$$R = 7 + 8 + 6 + 6 + 4 + 4 + 1 + 1 = 37$$
$$S = 74 - 45 = 29$$

d'où $\tau = 0.64$.

Les valeurs critiques de r_s et de τ au seuil 5 % bilatéral sont :

$$r_s = \pm 0.648 \qquad \text{et} \qquad \tau = \pm 1.96\sqrt{\frac{50}{90 \times 9}} = \pm 0.49$$

Les deux valeurs de τ et de r_s laissent donc apparaître une liaison significative entre les deux classements.

A part le cas où les variables sont ordinales, les coefficients de corrélation des rangs sont très utiles pour tester l'indépendance de deux variables ***non normales*** lorsque l'échantillon est ***petit*** : on sait en effet qu'on ne peut appliquer alors le test du coefficient de corrélation linéaire. Les tests de corrélation des rangs sont alors les seuls applicables, car ils ne dépendent pas de la distribution sous-jacente.

Ils sont robustes car insensibles à des valeurs aberrantes.

Les coefficients de corrélation de rangs sont en fait des ***coefficients de dépendance monotone*** car ils sont invariants pour toute transformation monotone croissante des variables.

Les coefficients de corrélation de rang permettent de tester l'existence d'une relation monotone entre deux variables. Ainsi le nuage de points suivant où $y = \ln(x)$ donne un coefficient de corrélation linéaire $r = 0.85$ mais des coefficients de Spearman et de Kendall égaux à 1.

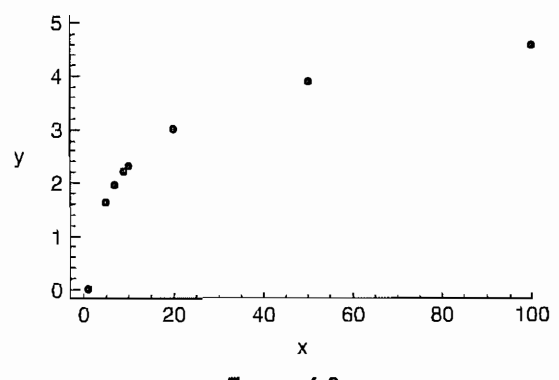


FIGURE 6.8

Lorsque les coefficients de corrélation de rang sont nettement supérieurs au coefficient de corrélation linéaire, des transformations monotones non linéaires sur certaines variables peuvent se révéler utiles.

6.3.3 Coefficients de Daniels et de Guttman

Les trois coefficients de corrélation (Pearson, Spearman, Kendall) peuvent être présentés comme 3 cas particuliers d'une même formule, dite formule de Daniels.

On considère pour toute paire d'individus i, j deux indices a_{ij} et b_{ij} le premier associé à la variable X, le deuxième associé à la variable Y (par exemple $a_{ij} = x_i - x_j$) et on définit le coefficient suivant :

$$\frac{\Sigma\Sigma a_{ij}b_{ij}}{\sqrt{(\Sigma\Sigma a_{ij}^2)(\Sigma\Sigma b_{ij}^2)}}$$

qui varie entre -1 et $+1$ d'après l'inégalité de Schwarz.

En prenant $a_{ij} = x_i - x_j$ et $b_{ij} = y_i - y_j$ on trouve le coefficient r de Bravais-Pearson ($\Sigma\Sigma(x_i - x_j)^2 = 2n^2 s_x^2$ par un calcul évident).

En prenant $a_{ij} = r_i - r_j$ et $b_{ij} = s_i - s_j$ où les r et les s sont les rangs de classement selon X et Y on obtient le coefficient de Spearman.

En prenant :

$$a_{ij} = \text{signe de } (x_i - x_j) = \frac{x_i - x_j}{|x_i - x_j|}$$
$$b_{ij} = \text{signe de } (y_i - y_j)$$

on obtient le coefficient τ de Kendall.

Mentionnons enfin le coefficient de monotonicité de Guttman :

$$\mu = \frac{\Sigma(x_i - x_j)(y_i - y_j)}{\Sigma|x_i - x_j||y_i - y_j|}$$

qui ne rentre pas dans la catégorie des coefficients de Daniels mais qui possède des propriétés intéressantes.

6.3.4 Le coefficient W de Kendall de concordance de p classements

Soient n individus (ou objets) été classés selon p critères (tableau 6.1) :

TABLEAU 6.1

Critères \ Objets	1	2		n	
1	r_{11}	r_{21}		r_{n1}	
2	r_{12}	r_{22}		r_{n2}	
p	r_{1p}	r_{2p}		r_{np}	
Total	$r_{1.}$	$r_{2.}$		$r_{n.}$	$r_{..}$

Chaque ligne du tableau est une permutation des entiers de 1 à n. La somme des termes d'une ligne étant $\frac{n(n+1)}{2}$, on a $r_{..} = p\frac{n(n+1)}{2}$.

Si les p classements étaient identiques (concordance parfaite) les totaux de colonnes $r_1, r_2, \ldots, r_n$ seraient égaux, à une permutation près, à $p, 2p, 3p, \ldots, np$; en effet, tous les termes d'une même colonne seraient identiques.

Pour étudier la concordance entre ces classements on utilise la statistique :

$$S = \sum_{i=1}^{n}\left(r_{i.} - \frac{r_{..}}{n}\right)^2$$

qui mesure la dispersion des totaux de colonnes par rapport à leur moyenne. On vérifie sans peine que S est maximal s'il y a concordance parfaite et que :

$$S_{\max} = \frac{p^2(n^3 - n)}{12}$$

Le coefficient de concordance de Kendall est :

$$W = \frac{12S}{p^2(n^3 - n)}$$

On a donc $0 \leq W \leq 1$.

Le cas limite $W = 0$ s'obtient si tous les totaux de colonnes sont identiques, une faible valeur de W indiquant l'indépendance entre les classements. On notera que la discordance parfaite entre p classements ne peut exister : il ne peut y avoir discordance parfaite entre plus de deux classements.

Le coefficient W est relié aux coefficients de corrélation des rangs de Spearman entre les classements pris deux à deux par la formule suivante :

$$\bar{r}_s = \frac{pW - 1}{p - 1}$$

où $\bar{r}_s$ est la moyenne arithmétique des C_p^2 coefficients de corrélation de Spearman entre classements.

Test de l'hypothèse H_0 d'indépendance mutuelle des p classements :

Pour les faibles valeurs de p et n, la distribution de W a pu être calculée sous l'hypothèse H_0 en considérant les $(n!)^p$ permutations équiprobables des p lignes du tableau.

On rejette H_0 si W est trop grand et on se reporte à la table fournie en annexe pour les valeurs critiques de S à $\alpha = 0.05$.

Pour $n \geq 15$ et pour $p < 7$, $\frac{(p-1)W}{1-W}$ est distribué sous H_0, comme une variable $F\left(n - 1 - \frac{2}{p} ; (p-1)\left(n - 1 - \frac{2}{p}\right)\right)$.

Pour $p \geq 7$ on admet que $p(n-1)\,W$ est distribué comme un χ^2_{n-1}.

Si l'on rejette l'hypothèse H_0 d'indépendance des p classements, quel classement final attribuer aux n objets ?

On admet en général la procédure suivante qui est de classer les objets selon l'ordre défini par la somme des colonnes ; cette procédure possède la propriété de maximiser la somme des coefficients de corrélation de Spearman entre le nouveau classement et les p classements initiaux[3].

Cas des *ex aequo* : pour calculer S, on remplace le rang des *ex aequo* dans un même classement par la moyenne arithmétique des rangs qu'ils auraient obtenus si il n'y avait pas eu d'*ex aequo* (ceci conserve la somme des lignes).

La valeur de S_{max} étant alors modifiée, on remplace W par :

$$W = \frac{12S}{p^2(n^3 - n) - p\sum_{j=1}^{p}(t_j^3 - t_j)}$$

où t_j est le nombre d'*ex aequo* du $j^{\text{ème}}$ classement.

6.4 LIAISON ENTRE UNE VARIABLE NUMÉRIQUE ET UNE VARIABLE QUALITATIVE

6.4.1 Le rapport de corrélation théorique (rappel)

La mesure, ici non symétrique, de la liaison est le rapport de corrélation $\eta_{Y/\mathscr{X}}$ défini par :

$$\eta^2_{Y/\mathscr{X}} = \frac{V[E(Y/\mathscr{X})]}{V(Y)}$$

En effet on peut appliquer η^2 lorsque la variable $\mathscr{X}$ n'est pas quantitative mais qualitative à k modalités (voir chapitre 3).

6.4.2 Le rapport de corrélation empirique

Si $\mathscr{X}$ a k catégories on notera $n_1, n_2, \ldots, n_k$ les effectifs observés et $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ les moyennes de Y pour chaque catégorie (il est indispensable qu'au moins un des n_i soit supérieurs à 1) et $\bar{y}$ la moyenne totale.

3 D'autres procédures basées sur la règle de la majorité de Condorcet sont possibles (voir l'ouvrage de J. F. Marcotorchino et P. Michaud, 1979) : recherche de l'ordre maximisant la somme des coefficients de Kendall.

Si l'on note e^2 l'équivalent empirique de η^2 on a :

$$e^2 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{s_y^2}$$

$e^2 = 0$ si $\bar{y}_1 = \bar{y}_2 = \ldots = \bar{y}_k$ d'où absence de dépendance en moyenne.

$e^2 = 1$ si tous les individus d'une catégorie de $\mathscr{X}$ ont même valeur de Y et ceci pour chaque catégorie car :

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{k} n_i s_i^2$$

où les s_i^2 sont les variances de Y à l'intérieur de chaque catégorie :

a) $\frac{1}{n}\sum n_i(\bar{y}_i - \bar{y})^2$ est appelée variance intercatégories.

b) $\frac{1}{n}\sum n_i s_i^2$ est appelée variance intracatégories.

On remarquera que si l'on attribue à chaque catégorie i de $\mathscr{X}$ une valeur numérique égale à $\bar{y}_i$ ce qui revient à transformer $\mathscr{X}$ en une variable numérique $\tilde{X}$ à k valeurs, e^2 est alors égal à $r^2(Y\,;\tilde{X})$ et que toute autre quantification conduit à une valeur de r^2 inférieure (voir plus loin).

Lorsqu'il n'y a que deux classes de moyennes $\bar{y}_1$ et $\bar{y}_2$:

$$e^2 = \frac{\frac{n_1 n_2}{n^2}(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

Pour déterminer à partir de quelle valeur e^2 est significatif on compare la variance inter à la variance intra : on montrera plus tard au chapitre 16 que si $\eta^2 = 0$ alors :

$$\frac{\dfrac{e^2}{k-1}}{\dfrac{1-e^2}{n-k}} = F(k-1\,;n-k)$$

Ce résultat suppose que les distributions conditionnelles de Y pour chaque catégorie de X sont gaussiennes avec même espérance et **même écart-type**.

On remarque que le nombre de classes intervient dans les degrés de liberté de la loi de Fisher-Snedecor : on ne peut donc comparer les rapports de corrélation entre Y et deux variables qualitatives ayant des nombres différents de catégories.

Lorsqu'aucune confusion n'est à craindre, l'usage est de noter η^2 le carré du rapport de corrélation empirique, c'est ce que nous ferons désormais.

Reprenons l'exemple du 5.3.2.2 sur les variations du taux de taxe d'habitation Y selon la zone géographique X : le rapport de corrélation est tel que :

$\eta^2(Y/X) = 0.56$ et correspond à $F = 20.05$

6.4.3 Interprétation géométrique et lien avec le coefficient de corrélation multiple

Associons à la variable qualitative $\mathscr{X}$ à k modalités les k variables numériques suivantes indicatrices des modalités :

$$\mathbb{1}^1 ; \mathbb{1}^2 ; \ldots ; \mathbb{1}^k$$

telles que :

$\mathbb{1}_i^j = 1$ si l'individu i est dans la catégorie j de $\mathscr{X}$;
$= 0$ sinon.

Soit alors X le tableau de données à n lignes et k colonnes correspondant aux indicatrices de $\mathscr{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ . & & \\ . & & \\ . & & \\ 0 & 0 & 1 \end{bmatrix}$$

Le total des éléments de la colonne j de $\mathbf{X}$ vaut n_j.

Un simple calcul permet alors de vérifier que :

$$\boxed{\eta^2_{\mathbf{Y}/\mathscr{X}} = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}}$$

si $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$ est centrée.

$\eta^2_{\mathbf{Y}/\mathscr{X}}$ est alors le cosinus carré de l'angle formé par le vecteur $\mathbf{y}$ centré et le sous-espace W de dimension k de $\mathbb{R}^n$ engendré par les variables indicatrices.

Le rapport de corrélation de $\mathbf{Y}$ en $\mathscr{X}$ s'identifie donc au coefficient de corrélation multiple avec les indicatrices de $\mathscr{X}$:

$$\boxed{\eta^2_{\mathbf{Y}/\mathscr{X}} = R^2(\mathbf{y} ; \mathbb{1}^1, \mathbb{1}^2, \ldots, \mathbb{1}^k)}$$

Définir une combinaison linéaire des indicatrices $\Sigma a_j \mathbb{1}^j$ revient à attribuer à chaque catégorie j une valeur numérique a_j, donc à rendre $\mathcal{X}$ numérique ce qui implique que :

$$\eta^2_{Y/\mathcal{X}} = r^2\left(\mathbf{y} \,;\, \sum_{j=1}^{k} \bar{y}_j \mathbb{1}^j\right) = \sup_{a_j} r^2\left(\mathbf{y} \,;\, \sum_{j=1}^{k} a_j \mathbb{1}^j\right)$$

6.5 LIAISON ENTRE DEUX VARIABLES QUALITATIVES

6.5.1 Tableau de contingence, marges et profils

Soit $\mathcal{X}$ et $\mathcal{Y}$ deux variables qualitatives à r et s catégories respectivement décrivant un ensemble de n individus. On présente usuellement les données sous la forme d'un tableau croisé appelé tableau de contingence à r lignes et s colonnes renfermant les effectifs n_{ij} d'individus tels que $\mathcal{X} = x_i$ et $\mathcal{Y} = y_j$ (voir tableau 6.2) :

TABLEAU 6.2

$\mathcal{Y}$ / $\mathcal{X}$	y_1	y_2	....	y_j	....	y_s	
x_1	n_{11}	n_{12}				n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}				n_{2s}	$n_{2.}$
⋮							
x_i				n_{ij}			$n_{i.}$
⋮							
x_r	n_{r1}	n_{r2}				n_{rs}	$n_{r.}$
	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.s}$	n

Avec des notations standard on a $n_{i.} = \sum_j n_{ij}$ et $n_{.j} = \sum_i n_{ij}$.

Les $n_{i.}$ et les $n_{.j}$ s'appellent respectivement ***marges*** en lignes et ***marges*** en colonnes.

La constitution d'un tel tableau est l'opération que les praticiens des enquêtes appellent un « tri croisé ».

Deux lectures différentes d'un même tableau de contingence sont possibles selon que l'on privilégie l'une ou l'autre des deux variables : lecture en ligne ou lecture en colonnes.

On appelle tableau des ***profils-lignes*** le tableau des fréquences conditionnelles $\dfrac{n_{ij}}{n_{i.}}$ (la somme de chaque ligne est ramenée à 100 %) et tableau des ***profils-colonnes*** le tableau des fréquences conditionnelles $\dfrac{n_{ij}}{n_{.j}}$ (le total de chaque colonne est alors ramené à 100 %).

Exemple : Le tableau 6.3 provient de l'enquête sur les vacances des Français en 1999, publiée par l'INSEE en mai 2002.

On appelle vacances tout déplacement comportant au moins 4 nuitées consécutives en dehors du domicile, effectué pour des motifs autres que professionnels, études ou santé. Un voyage peut comporter un ou plusieurs séjours (4 nuits consécutives au même endroit).

En 1999 près d'un français sur quatre n'était pas parti en vacances, le tableau de contingence ne concerne donc que ceux qui sont partis.

L'unité statistique est ici le séjour, décrit par deux variables qualitatives : la catégorie socio-professionnelle de la personne de référence du ménage en 8 modalités et le mode d'hébergement en 9 modalités. La taille de l'échantillon est 18 352.

TABLEAU 6.3 Tableau de contingence

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV
Agriculteurs	41	47	13	59	17	26	4	9	19
Artisans, commerçants, chefs d'entreprise	220	260	71	299	120	42	64	35	29
Cadres et professions intellectuelles supérieures	685	775	450	1242	706	139	122	100	130
Professions intermédiaires	485	639	292	1250	398	189	273	68	193
Employés	190	352	67	813	163	92	161	49	72
Ouvriers	224	591	147	1204	181	227	306	74	114
Retraités	754	393	692	1158	223	25	195	47	115
Autres inactifs	31	34	2	225	42	33	5	6	14

On déduit du tableau 6.3 les deux tableaux de profils suivants (6.4 et 6.5) qui permettent deux types de lecture : le tableau des profils-lignes permet de comparer les modes d'hébergement des différentes catégories socio-professionnelles (où vont les cadres ? etc.) tandis que le tableau des profils-colonnes permet de savoir qui fréquente tel mode (qui sont les clients des hôtels ?).

Tableau 6.4 Tableau des profils-lignes

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV	Total
Agriculteurs	0.174	0.200	0.055	0.251	0.072	0.111	0.017	0.038	0.081	1
Artisans, commerçants, chefs d'entreprise	0.193	0.228	0.062	0.262	0.105	0.037	0.056	0.031	0.025	1
Cadres et professions intellectuelles supérieures	0.158	0.178	0.103	0.286	0.162	0.032	0.028	0.023	0.030	1
Professions intermédiaires	0.128	0.169	0.077	0.330	0.105	0.050	0.072	0.018	0.051	1
Employés	0.097	0.180	0.034	0.415	0.083	0.047	0.082	0.025	0.037	1
Ouvriers	0.073	0.193	0.048	0.392	0.059	0.074	0.100	0.024	0.037	1
Retraités	0.209	0.109	0.192	0.321	0.062	0.007	0.054	0.013	0.032	1
Autres inactifs	0.079	0.087	0.005	0.574	0.107	0.084	0.013	0.015	0.036	1

Tableau 6.5 Tableau des profils colonnes

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV
Agriculteurs	0.016	0.015	0.007	0.009	0.009	0.034	0.004	0.023	0.028
Artisans, commerçants, chefs d'entreprise	0.084	0.084	0.041	0.048	0.065	0.054	0.057	0.090	0.042
Cadres et professions intellectuelles supérieures	0.260	0.251	0.260	0.199	0.38`2	0.180	0.108	0.258	0.190
Professions intermédiaires	0.184	0.207	0.168	0.200	0.215	0.245	0.242	0.175	0.281
Employés	0.072	0.114	0.039	0.130	0.088	0.119	0.142	0.126	0.105
Ouvriers	0.085	0.191	0.085	0.193	0.098	0.294	0.271	0.191	0.166
Retraités	0.287	0.127	0.399	0.185	0.121	0.032	0.173	0.121	0.168
Autres inactifs	0.012	0.011	0.001	0.036	0.023	0.043	0.004	0.015	0.020
Total	1	1	1	1	1	1	1	1	1

On remarquera que la moyenne des profils-lignes (avec des poids correspondant aux effectifs marginaux des lignes) n'est autre que le profil marginal des colonnes :

$$\sum_{i=1}^{r} \frac{n_{ij}}{n_{i.}}\left(\frac{n_{i.}}{n}\right) = \frac{n_{.j}}{n}$$

et que l'on a de même :
$$\sum_{j=1}^{s} \frac{n_{ij}}{n_{.j}}\left(\frac{n_{.j}}{n}\right) = \frac{n_{i.}}{n}$$

6.5.2 L'écart à l'indépendance

Lorsque tous les profils-lignes sont identiques on peut parler d'indépendance entre $\mathscr{X}$ et $\mathscr{Y}$ puisque la connaissance de $\mathscr{X}$ ne change pas les distributions conditionnelles de $\mathscr{Y}$. Il s'ensuit d'ailleurs que tous les profils-colonnes sont également identiques.

On doit donc avoir $\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \ldots = \frac{n_{rj}}{n_{r.}}$ $\forall j$, ce qui entraîne $\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$ par sommation des numérateurs et dénominateurs.

L'indépendance empirique se traduit donc par $n_{ij} = \frac{n_{i.}n_{.j}}{n}$.

6.5.2.1 Le χ^2 d'écart à l'indépendance et les autres mesures associées

On adopte généralement la mesure suivante de liaison d^2 notée aussi X^2 ou χ^2 (voir plus loin) :

$$\boxed{d^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}}$$

On voit que d^2 est nul dans le cas de l'indépendance. Quelle est sa borne supérieure et dans quel cas est-elle atteinte ? Il faut pour cela utiliser le résultat suivant obtenu par développement du carré :

$$d^2 = n\left[\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

Comme : $\frac{n_{ij}}{n_{.j}} \leq 1$ on a : $\frac{n_{ij}^2}{n_{i.}n_{.j}} \leq \frac{n_{ij}}{n_{.j}}$

D'où :
$$\sum_i \sum_j \frac{n_{ij}^2}{n_{i.}n_{.j}} \leq \sum_i \sum_j \frac{n_{ij}}{n_{.j}} = \sum_j \frac{\sum_i n_{ij}}{n_{.j}} = \sum_{j=1}^{s} \frac{n_{.j}}{n_{.j}} = s$$

D'où $d^2 \leq n(s-1)$. On pourrait montrer de même que $d^2 \leq n(r-1)$. On a donc :

$$\boxed{\frac{d^2}{n} \leq \inf(s-1\,;r-1)}$$

La borne étant atteinte dans le cas de la dépendance fonctionnelle. En effet $d^2 = n(s-1)$ si $\frac{n_{ij}}{n_{i.}} = 1\ \forall i$, c'est-à-dire s'il n'existe qu'une case non nulle dans chaque ligne.

Ce cas est celui où $\mathcal{Y}$ est fonctionnellement lié à $\mathcal{X}$: ce qui n'implique pas que $\mathcal{X}$ soit lié fonctionnellement à $\mathcal{Y}$ comme on le voit sur la figure 6.9.

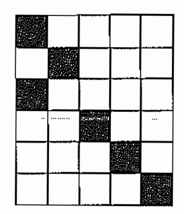


FIGURE 6.9

Le cas de la dépendance fonctionnelle réciproque nécessite $r = s$: après permutation des lignes ou des colonnes le tableau de contingence peut alors se mettre sous forme diagonale.

Divers coefficients liés au d^2 ont été proposés pour obtenir une mesure comprise entre 0 (indépendance) et 1 (liaison fonctionnelle). Citons :

- le coefficient de contingence de K. Pearson $\left(\frac{d^2}{n+d^2}\right)^{1/2} = C$;
- le coefficient de Tschuprow $\left(\frac{d^2}{n\sqrt{(r-1)(s-1)}}\right)^{1/2} = T$;
- le coefficient de Cramer $\left(\frac{d^2}{n\inf\{(s-1)\,;(r-1)\}}\right)^{1/2}$

d^2/n est usuellement noté φ^2. Pour l'exemple des vacances présenté plus haut on a :

$$d^2 = 1989 \qquad C = 0.31$$
$$T = 0.12 \qquad V = 0.12$$

La construction du tableau des $\frac{n_{i.}n_{.j}}{n}$ (tableau d'indépendance) et sa comparaison avec le tableau des n_{ij} est en général instructive : en particulier le calcul pour chaque case du terme :

$$\frac{(n_{ij} - n_{i.}n_{.j})^2}{\frac{n_{i.}n_{.j}}{n}} \cdot \frac{1}{d^2}$$

appelé ***contribution au χ^2*** permet de mettre en évidence les associations significatives entre catégories des deux variables. Le signe de la différence $n_{ij} - \frac{n_{i.} n_{.j}}{n}$ indique alors s'il y a association positive ou négative entre les catégories i de $\mathcal{X}$ et j de $\mathcal{Y}$.

Un tel calcul devrait être systématiquement associé à chaque calcul de χ^2.

On remarque que les marges des tableaux (n_{ij}) et $\left(\frac{n_{i.} n_{.j}}{n}\right)$ étant les mêmes par construction il suffit de calculer $(r-1)(s-1)$ (le degré de liberté) termes du tableau d'indépendance et de déduire les autres par différence.

Le tableau 6.6 donne pour chaque case l'effectif théorique et le χ^2 correspondant. Comme il y a 72 cases, le χ^2 moyen par case est de 27.6 : on a mis en grisé les cases où le χ^2 dépasse 60 : ce sont les cases où il existe une sur- ou une sous-représentation importante par rapport à une répartition « au hasard ».

TABLEAU 6.6

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV
Agriculteurs	33.35 1.75	39.2 1.55	21.99 3.67	79.25 5.18	23.46 1.78	9.8 26.77	14.33 7.45	4.92 3.38	8.7 12.2
Artisans, commerçants, chefs d'entreprise	161.79 20.95	190.14 25.66	10.67 11.93	384.47 19	113.8 0.34	47.55 0.65	69.51 0.44	23.87 5.19	42.2 4.13
Cadres et professions intellectuelles supérieures	617.2 7.45	725.8 3.39	406.93 4.56	1466.72 34.43	434.15 170.22	181.4 9.91	265.18 77.31	91.05 0.88	160.99 5.96
Professions intermédiaires	537.44 5.12	631.64 0.09	354.34 10.97	1277.18 0.58	378.05 1.05	157.96 6.10	230.91 7.67	79.29 1.61	140.18 19.9
Employés	278.01 27.86	326.75 1.95	183.3 73.79	660.68 35.12	195.56 5.42	81.71 1.3	119.45 14.45	41.02 1.55	75.52 0.00
Ouvriers	435.4 102.64	511.72 12.28	287.07 68.34	1034.7 27.70	306.7 51.24	127.97 76.63	187.07 75.6	64.23 1.48	113.57 0.00
Retraités	511.18 115.34	600.79 71.86	337.03 373.86	1214.79 2.65	359.68 51.88	150.25 104.41	219.63 2.76	75.41 10.71	133.34 2.52
Autres inactifs	55.63 10.91	65.38 15.06	36.68 32.79	132.2 65.14	39.13 0.21	16.35 16.95	23.9 14.95	8.21 0.59	14.51 0.02

L'analyse des correspondances étudiée plus loin permet une représentation graphique des écarts à l'indépendance : on y retrouvera ces phénomènes.

6.5.2.2 Cas des tableaux 2 × 2

Si $\mathcal{X}$ et $\mathcal{Y}$ n'ont que deux modalités chacune le tableau de contingence (tableau 6.7) n'a alors que 4 cases d'effectifs *abcd*.

TABLEAU 6.7

$\mathcal{X}$ \ $\mathcal{Y}$	1	2
1	a	b
2	c	d

d^2 peut alors s'exprimer par la formule :

$$d^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

Si l'on calcule le coefficient de corrélation linéaire entre $\mathcal{X}$ et $\mathcal{Y}$ en attribuant des valeurs arbitraires à leurs catégories (par exemple 1 et 2 mais toutes les quantifications sont ici équivalentes) on montre que $\varphi^2 = r^2$.

Remarquons que dans le cas des variables dichotomiques la non corrélation entraîne l'indépendance.

6.5.2.3 Caractère significatif de l'écart à l'indépendance

A partir de quelle valeur peut-on considérer que la liaison est significative ? En anticipant sur la théorie des tests exposée au chapitre 14 voici la démarche : si les n observations étaient prélevées dans une population où $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes ($p_{ij} = p_{i.}\,p_{.j}$) quelles seraient les valeurs probables de d^2 ?

En s'appuyant sur les résultats du chapitre 4 paragraphe 4.4, on montre qu'alors d^2 est une réalisation d'une variable aléatoire D^2 suivant approximativement une loi $\chi^2_{(r-1)(s-1)}$; en effet les $n_{ij} - \dfrac{n_{i.}n_{.j}}{n}$ sont liés par $(r - 1)(s - 1)$ relations linéaires puisque les marges sont communes aux deux tableaux (ou encore en d'autres termes puisqu'on estime les $p_{i.}$ par $\dfrac{n_{i.}}{n}$ et les $p_{.j}$ par $\left.\dfrac{n_{.j}}{n}\right)$.

Il suffit alors de se fixer un risque d'erreur α, c'est-à-dire une valeur qui, s'il y avait indépendance, n'aurait qu'une probabilité faible d'être dépassée (on prend usuellement $\alpha = 5\ \%$ ou $1\ \%$).

On rejettera donc l'hypothèse d'indépendance si d^2 est supérieur à la valeur critique qu'une variable $x^2_{(r-1)(s-1)}$ a une probabilité a de dépasser.

Ainsi sur l'exemple : le degré de liberté du χ^2 est $(9-1)(8-1) = 56$. La valeur de d^2 est très élevée : $d^2 = 1989$.

La valeur critique à 1 % d'un χ^2_{56} est 83.5.

On doit donc rejeter l'hypothèse d'indépendance entre catégorie professionnelle et mode d'hébergement.

Pour les tableaux 2×2 où le degré de liberté vaut 1 on recommande généralement d'effectuer la correction de Yates :

$$d^2 = \frac{n\left[|ad - bc| - \dfrac{n}{2}\right]^2}{(a+b)(a+c)(b+d)(c+d)}$$

L'espérance d'un χ^2 étant égale à son degré de liberté on voit que d^2 est d'autant plus grand que le nombre de catégories des deux variables est élevé. On ne peut donc comparer des d^2 correspondant à des tableaux de tailles différentes pour une même valeur de n : un d^2 de 4 pour un tableau 2×4 ne révèle pas une dépendance plus forte qu'un d^2 de 2.7 pour un tableau 2×2 bien au contraire : afin de comparer ce qui est comparable et de s'affranchir du problème des degrés de liberté il vaut mieux utiliser comme indice de liaison la probabilité $P(\chi^2 < d^2)$. On trouve ainsi :

$$P(\chi^2_1 < 2.7) = 0.9 \qquad \text{et} \qquad P(\chi^2_3 < 4) \simeq 0.75$$

6.5.2.4 *Autres mesures de dépendance*

Les indices dérivés du χ^2 sont loin d'être les seules mesures de dépendance utilisables, elles ont d'ailleurs été souvent critiquées. La littérature statistique abonde en la matière et le problème est d'ailleurs celui du trop grand nombre d'indices proposés. On se reportera utilement aux ouvrages de Goodman et Kruskal et de Marcotorchino (1979).

Signalons toutefois pour son intérêt théorique le G^2 ou khi-deux de vraisemblance :

$$G^2 = 2\sum_i \sum_j n_{ij} \ln\left(\frac{n_{ij}}{\dfrac{n_{i.}n_{.j}}{n}}\right)$$

qui sous l'hypothèse d'indépendance suit une loi $\chi^2_{(r-1)(s-1)}$.

6.5.3 Un indice non symétrique de dépendance : le τ_b de Goodman et Kruskal

$$\tau_{by/x} = \frac{\displaystyle\sum_i \sum_j \frac{n_{ij}^2}{n n_{i.}} - \sum_j \left(\frac{n_{.j}}{n}\right)^2}{1 - \displaystyle\sum_j \left(\frac{n_{.j}}{n}\right)^2}$$

Cet indice résulte du raisonnement suivant : si l'on ignore $\mathscr{X}$, la probabilité (estimée) qu'une observation appartienne à la catégorie j de $\mathscr{Y}$ est $\frac{n_{.j}}{n}$: en affectant aléatoirement cette observation selon les probabilités $\frac{n_{.j}}{n}$ on a alors une proportion estimée de classements corrects égale à $\sum_j \left(\frac{n_{.j}}{n}\right)^2$.

Si l'on connaît la catégorie i de $\mathscr{X}$ l'affectation se fait alors selon les fréquences conditionnelles $\frac{n_{ij}}{n_{i.}}$ d'où une proportion estimée de classements corrects égale à $\sum_i \sum_j \frac{n_{ij}}{n_{i.}} \frac{n_{ij}}{n}$.

Le τ de Goodman et Kruskal mesure donc le taux de décroissance du pourcentage de prédictions incorrectes.

On a par définition $0 \leq \tau_b \leq 1$ avec $\tau_b = 0$ dans le cas de l'indépendance et $\tau_b = 1$ pour la liaison fonctionnelle.

En introduisant les tableaux de variables indicatrices $\mathbf{X}_1$ et $\mathbf{X}_2$ associées aux deux variables $\mathscr{X}$ et $\mathscr{Y}$ on trouve :

$$\tau = \frac{\text{Trace}\,(\mathbf{X}'_2 \mathbf{A}_1^0 \mathbf{X}_2)}{\text{Trace}\,\mathbf{V}_{22}}$$

où $\mathbf{A}_1^0$ est le projecteur sur l'espace des combinaisons linéaires de ***moyenne nulle*** des indicatrices de $\mathbf{X}_1$.

τ n'est autre que le coefficient de redondance $R^2(\mathbf{X}_2 : \mathbf{X}_1)$ de Stewart et Love (voir chapitre 8).

6.5.4 Le kappa de Cohen

Ce coefficient est destiné à mesurer l'accord entre deux variables qualitatives ayant les mêmes modalités dans le contexte suivant : n unités statistiques sont réparties selon p catégories par deux observateurs. Si les deux observateurs concordent parfaitement, le tableau de contingence doit être diagonal : seuls les effectifs n_{ii} sont non-nuls.

La proportion d'accords observés est $P_0 = \frac{1}{n}\sum_{i=1}^{p} n_{ii}$.

Si les deux variables étaient indépendantes la probabilité d'être dans l'une quelconque des cases diagonales serait $\sum_{i=1}^{p} p_{i.}\, p_{.i}$ que l'on estime par $P_e = \frac{1}{n^2}\sum_{i=1}^{p} n_{i.}\, n_{.i}$ appelé pourcentage d'accords aléatoires.

Le coefficient kappa s'écrit alors :

$$\kappa = \frac{P_0 - P_e}{1 - P_e} = \frac{\frac{1}{n}\sum_{i=1}^{p} n_{ii} - \frac{1}{n^2}\sum_{i=1}^{p} n_{i.}\, n_{.i}}{1 - \frac{1}{n^2}\sum_{i=1}^{p} n_{i.}\, n_{.i}}$$

kappa est compris entre -1 et $+1$ (accord maximal).

7 L'analyse en composantes principales

Dans la plupart des applications on observe non pas une variable par individu, mais un nombre p souvent élevé. L'étude séparée de chacune de ces variables et celles des couples selon les techniques exposées précédemment est une phase indispensable dans le processus de dépouillement des données mais tout à fait insuffisante.

Il faut donc analyser les données en tenant compte de leur caractère multidimensionnel ; l'analyse en composantes principales est alors une méthode particulièrement puissante pour explorer la structure de telles données. C'est également la « mère » de la plupart des méthodes descriptives multidimensionnelles.

7.1 TABLEAUX DE DONNÉES, RÉSUMÉS NUMÉRIQUES ET ESPACES ASSOCIÉS

7.1.1 Les données et leurs caractéristiques

7.1.1.1 Le tableau des données

Les observations de p variables sur n individus sont rassemblées en un tableau rectangulaire X à n lignes et p colonnes :

$$\mathbf{X} = \begin{array}{c} 1 \\ 2 \\ . \\ i \\ . \\ . \\ . \\ n \end{array} \overset{\begin{array}{cccccc} 1 & 2 & \dots & j & \dots & p \end{array}}{\begin{bmatrix} & & & . & & \\ & & & . & & \\ & & & . & & \\ . & . & \dots & x_i^j & \dots & . \\ & & & . & & \\ & & & . & & \\ & & & . & & \\ & & & . & & \end{bmatrix}}$$

x_i^j est la valeur prise par la variable n^o j sur le $i^{\text{ème}}$ individu.

Dans une optique purement descriptive on identifiera une variable à la colonne de $\mathbf{X}$ correspondante : une variable n'est rien d'autre que la liste des n valeurs qu'elle prend sur les n individus :

$$\mathbf{x}^j = \begin{bmatrix} x_1^j \\ . \\ . \\ . \\ x_n^j \end{bmatrix}$$

On identifiera de même l'individu i au vecteur $\mathbf{e}_i$ à p composantes :

$$\mathbf{e}_i' = (x_i^1 \dots x_i^p)$$

7.1.1.2 Poids et centre de gravité

Si les données ont été recueillies à la suite d'un tirage aléatoire à probabilités égales, les n individus ont tous même importance, $1/n$, dans le calcul des caractéristiques de l'échantillon. Il n'en est pas toujours ainsi et il est utile pour certaines applications de travailler avec des poids p_i éventuellement différents d'un individu à l'autre (échantillons redressés ; données regroupées . . .).

Ces poids, qui sont des nombres positifs de somme 1 comparables à des fréquences, sont regroupés dans une matrice diagonale $\mathbf{D}$ de taille n :

$$\mathbf{D} = \begin{bmatrix} p_1 & & & & 0 \\ & p_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ 0 & & & & p_n \end{bmatrix}$$

Dans le cas le plus usuel de poids égaux, $\mathbf{D} = \dfrac{1}{n}\mathbf{I}$.

Le vecteur $\mathbf{g}$ des moyennes arithmétiques de chaque variable $\mathbf{g}' = (\bar{x}^1 ; \bar{x}^2 ; \dots ; \bar{x}^p)$ définit le point moyen, ou centre de gravité du nuage.

On a $\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1.

Le tableau $\mathbf{Y}$ tel que $y_i^j = x_i^j - \bar{x}^j$ est le tableau centré associé à $\mathbf{X}$.

On a $\mathbf{Y} = \mathbf{X} - \mathbf{1}\mathbf{g}' = (\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D})\mathbf{X}$.

7.1.1.3 Matrice de variance-covariance et matrice de corrélation

La formule établie au chapitre précédent avec des poids égaux à $1/n$ se généralise comme suit :

$$\boxed{\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D}\mathbf{Y}}$$

On a également : $$\mathbf{X'DX} = \sum_{i=1}^{n} p_i \mathbf{e}_i \mathbf{e}'_i$$

Cette dernière formule est utile pour les calculs numériques car elle ne suppose pas la mise en mémoire du tableau **X** mais seulement la lecture successive des données.

Si l'on note $\mathbf{D}_{1/s}$ la matrice diagonale des inverses des écarts-types :

$$\mathbf{D}_{1/s} = \begin{bmatrix} 1/s_1 & & & 0 \\ & . & & \\ & & . & \\ & & & . \\ 0 & & & 1/s_p \end{bmatrix}$$

et $\mathbf{D}_{1/s^2}$ la matrice diagonale des inverses des variances, le tableau des données centrées et réduites **Z** tel que :

$$z_i^j = \frac{x_i^j - \bar{x}^j}{s_j}$$

est donc : $$\mathbf{Z} = \mathbf{YD}_{1/s}$$

La matrice regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux est notée **R** :

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ . & 1 & & . \\ . & & & . \\ . & & & . \\ r_{p1} & & & 1 \end{bmatrix}$$

Rappelons que $\mathbf{R} = \mathbf{D}_{1/s}\mathbf{VD}_{1/s} = \mathbf{Z'DZ}$.

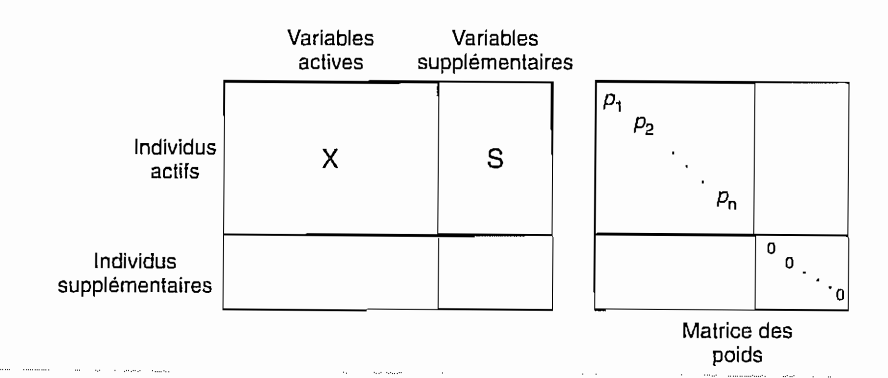
R est la matrice de variance-covariance des données centrées et réduites et résume la structure des dépendances linéaires entre les p variables prise 2 à 2.

7.1.1.4 Données actives et supplémentaires

Le tableau X ne représente souvent qu'une partie de l'information disponible, et cela pour diverses raisons : on ne s'intéresse qu'aux liaisons entre certaines variables, les variables qualitatives sont par nature exclues de l'analyse etc. Les variables disponibles se partagent donc en deux ensembles : les variables actives qui serviront au calcul des axes principaux et les variables supplémentaires, appelées également variables illustratives qui seront reliées *a posteriori* aux résultats de l'ACP.

On peut également n'utiliser qu'une partie des individus, soit pour valider les résultats, soit parce que certains n'auront leur données disponibles qu'ultérieurement, ou parce que

leurs données sont suspectes. Mettre des individus en supplémentaire revient à leur attribuer un poids nul.



7.1.2 L'espace des individus

Chaque individu étant un point défini par p coordonnées est considéré comme un élément d'un espace vectoriel F appelé l'espace des individus. L'ensemble des n individus est alors un « nuage » de points dans F et **g** en est le centre de gravité.

L'espace F est muni d'une structure euclidienne afin de pouvoir définir des distances entre individus.

7.1.2.1 Le rôle de la métrique

Comment mesurer la distance entre deux individus ? Cette question primordiale doit être résolue avant toute étude statistique car les résultats obtenus en dépendent dans une large mesure.

En physique, la distance entre deux points de l'espace se calcule facilement par la formule de Pythagore : le carré de la distance est la somme des carrés des différences des coordonnées, car les dimensions sont de même nature : ce sont des longueurs que l'on mesure avec la même unité :

$$d^2 = (x_1^k - x_2^k)^2 + (x_1^j - x_2^j)^2 + \cdots$$

Il n'en est pas de même en statistique où chaque dimension correspond à un caractère qui s'exprime avec son unité particulière : comment calculer la distance entre deux individus décrits par les trois caractères : âge, salaire, nombre d'enfants ?

La formule de Pythagore est alors aussi arbitraire qu'une autre. Si l'on veut donner des importances différentes à chaque caractère, pourquoi ne pas prendre une formule du type :

$$d^2 = a_1(x_1^1 - x_2^1)^2 + a_2(x_1^2 - x_2^2)^2 + \cdots + a_p(x_1^p - x_2^p)^2$$

ce qui revient à multiplier par $\sqrt{a_j}$ chaque caractère (on prendra bien sûr des a_j positifs).

De plus, la formule de Pythagore n'est valable que si les axes sont perpendiculaires, ce que l'on conçoit aisément dans l'espace physique. Mais en statistique ce n'est que par pure convention que l'on représente les caractères par des axes perpendiculaires : on aurait pu tout aussi bien prendre des axes obliques.

On utilisera donc la formulation générale suivante : la distance entre deux individus $\mathbf{e}_i$ et $\mathbf{e}_j$ est définie par la forme quadratique :

$$d^2(\mathbf{e}_i\,;\,\mathbf{e}_j) = (\mathbf{e}_i - \mathbf{e}_j)'\mathbf{M}(\mathbf{e}_i - \mathbf{e}_j)$$

où $\mathbf{M}$ est une matrice symétrique de taille p définie positive. L'espace des individus est donc muni du produit scalaire : $\langle \mathbf{e}_i\,;\,\mathbf{e}_j \rangle = \mathbf{e}_i'\mathbf{M}\mathbf{e}_j$

En théorie, le choix de la matrice $\mathbf{M}$ dépend de l'utilisateur qui seul peut préciser la métrique adéquate. En pratique les métriques usuelles en Analyse en Composantes Principales (ACP) sont en nombre réduit : à part la métrique $\mathbf{M} = \mathbf{I}$ qui revient à utiliser le produit scalaire usuel, la métrique la plus utilisée (et qui est souvent l'option par défaut des logiciels) est la métrique diagonale des inverses des variances :

$$\mathbf{M} = \mathbf{D}_{1/s^2} = \begin{bmatrix} 1/s_1^2 & & & & 0 \\ & 1/s_2^2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ 0 & & & & 1/s_p^2 \end{bmatrix}$$

ce qui revient à diviser chaque caractère par son écart-type : entre autres avantages, la distance entre deux individus ne dépend plus des unités de mesure puisque les nombres x_i^j/s_j sont sans dimension, ce qui est très utile lorsque les variables ne s'expriment pas avec les mêmes unités.

Surtout, cette métrique donne à chaque caractère la même importance quelle que soit sa dispersion ; l'utilisation de $\mathbf{M} = \mathbf{I}$ conduirait à privilégier les variables les plus dispersées, pour lesquelles les différences entre individus sont les plus fortes, et à négliger les différences entre les autres variables. La métrique $\mathbf{D}_{1/s^2}$ rétablit alors l'équilibre entre les variables en donnant à toutes la variance 1.

Nous avons vu qu'utiliser une métrique diagonale :

$$\mathbf{D}_a = \begin{bmatrix} a_1 & & & & \\ & a_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & a_p \end{bmatrix}$$

revient à multiplier les caractères par $\sqrt{a_i}$ et utiliser ensuite la métrique usuelle $\mathbf{M} = \mathbf{I}$. Ce résultat se généralise à une métrique $\mathbf{M}$ quelconque de la manière suivante :

On sait que toute matrice symétrique positive $\mathbf{M}$ peut s'écrire $\mathbf{M} = \mathbf{T}'\mathbf{T}$. Le produit scalaire entre deux individus avec la métrique $\mathbf{M}$ peut donc s'écrire :

$$\langle \mathbf{e}_1 \, ; \mathbf{e}_2 \rangle = \mathbf{e}_1' \, \mathbf{M}\mathbf{e}_2 = \mathbf{e}_1' \, \mathbf{T}'\mathbf{T}\mathbf{e}_2 = (\mathbf{T}\mathbf{e}_2)'\mathbf{T}\mathbf{e}_1$$

Tout se passe donc comme si l'on utilisait la métrique $\mathbf{I}$ sur des données transformées, c'est-à-dire sur le tableau $\mathbf{XT}'$.

7.1.2.2 L'inertie

On appelle inertie totale du nuage de points la moyenne pondérée des carrés des distances des points au centre de gravité :

$$I_g = \sum_{i=1}^{n} p_i(\mathbf{e}_i - \mathbf{g})'\mathbf{M}(\mathbf{e}_i - \mathbf{g}) = \sum_i p_i \, \|\mathbf{e}_i - \mathbf{g}\|^2$$

L'inertie en un point $\mathbf{a}$ quelconque est définie par :

$$I_a = \sum_i p_i(\mathbf{e}_i - \mathbf{a})'\mathbf{M}(\mathbf{e}_i - \mathbf{a})$$

On a la relation de Huyghens :

$$I_a = I_g + (\mathbf{g} - \mathbf{a})'\mathbf{M}(\mathbf{g} - \mathbf{a}) = I_g + \|\mathbf{g} - \mathbf{a}\|^2$$

Si $\mathbf{g} = \mathbf{0}$:

$$I_g = \sum_{i=1}^{n} p_i \mathbf{e}_i' \mathbf{M}\mathbf{e}_i$$

Par ailleurs, on démontre aisément que l'inertie totale vérifie la relation :

$$2I_g = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j (\mathbf{e}_i - \mathbf{e}_j)'\mathbf{M}(\mathbf{e}_i - \mathbf{e}_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \|\mathbf{e}_i - \mathbf{e}_j\|^2$$

soit la moyenne des carrés de toutes les distances entre les n individus.

L'inertie totale est la trace de la matrice $\mathbf{MV}$ (ou $\mathbf{VM}$) :

$$\boxed{I_g = \text{Trace } \mathbf{MV} = \text{Trace } \mathbf{VM}}$$

En effet, $p_i \mathbf{e}_i' \mathbf{M}\mathbf{e}_i$ étant un scalaire, grâce à la commutativité sous la trace :

$$I_g = \text{Trace}\left(\sum_{i=1}^{n} p_i \mathbf{e}_i' \mathbf{M}\mathbf{e}_i\right) = \text{Trace}\left(\sum_{i=1}^{n} \mathbf{M}\mathbf{e}_i p_i \mathbf{e}_i'\right)$$
$$= \text{Trace } \mathbf{MX'DX} \qquad = \text{Trace } \mathbf{MV}$$

- si $\mathbf{M} = \mathbf{I}$ l'inertie est égale à la somme des variances des p variables ;
- si $\mathbf{M} = \mathbf{D}_{1/s^2}$: Trace $\mathbf{MV}$ = Trace $(\mathbf{D}_{1/s^2}\mathbf{V})$ = Trace $(\mathbf{D}_{1/s}\mathbf{V}\mathbf{D}_{1/s})$, ce qui est égal à Trace $\mathbf{R} = p$. L'inertie est donc égale au nombre de variables et ne dépend pas de leurs valeurs.

7.1.3 L'espace des variables

Chaque variable x^j est en fait une liste de n valeurs numériques : on la considère comme un vecteur $\mathbf{x}^j$ d'un espace E à n dimensions appelé espace des variables.

7.1.3.1 La métrique des poids

Pour étudier la proximité des variables entre elles il faut munir cet espace d'une métrique, c'est-à-dire trouver une matrice d'ordre n définie positive symétrique. Ici il n'y a pas d'hésitation comme pour l'espace des individus et le choix se porte sur la matrice diagonale des poids $\mathbf{D}$ pour les raisons suivantes :

- Le produit scalaire de deux variables $\mathbf{x}^j$ et $\mathbf{x}^k$ qui vaut $\mathbf{x}'^j\mathbf{D}\mathbf{x}^k = \sum_{i=1}^{n} p_i x_i^k x_i^j$ n'est autre que la covariance s_{jk} si les deux variables sont centrées.
- La norme d'une variable $\|\mathbf{x}^j\|_{\mathbf{D}}$ est alors $\|\mathbf{x}^j\|_{\mathbf{D}}^2 = s_j^2$; en d'autres termes la « longueur » d'une variable est égale à son écart-type.
- L'angle θ_{jk} entre deux variables centrées est donné par :

$$\cos\theta_{jk} = \frac{\langle \mathbf{x}^j ; \mathbf{x}^k \rangle}{\|\mathbf{x}^j\| \, \|\mathbf{x}^k\|} = \frac{s_{jk}}{s_j s_k}$$

Le cosinus de l'angle entre deux variables centrées n'est autre que leur coefficient de corrélation linéaire (chapitre 6).

Si dans l'espace des individus on s'intéresse aux distances entre points, dans l'espace des variables on s'intéressera plutôt aux angles en raison de la propriété précédente.

7.1.3.2 Variables engendrées par un tableau de données

A une variable $\mathbf{x}^j$ on peut associer un axe de l'espace des individus F et un vecteur de l'espace des variables E.

On peut également déduire de $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^p$ de nouvelles variables par combinaison linéaire, ce qui revient à projeter les individus sur de nouveaux axes de F.

Considérons un axe Δ de l'espace des individus engendré par un vecteur unitaire $\mathbf{a}$ (c'est-à-dire de $\mathbf{M}$-norme 1) et projetons les individus sur cet axe (projection $\mathbf{M}$-orthogonale) (fig. 7.1).

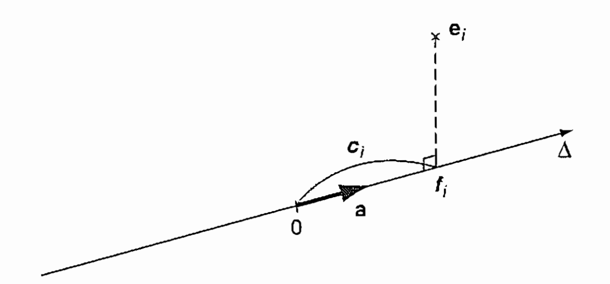


FIGURE 7.1

La liste des coordonnées c_i des individus sur Δ forme une nouvelle variable ou composante, **c**. Comme $c_i = \mathbf{a'Me}_i = \mathbf{e}'_i\mathbf{Ma} = \langle \mathbf{a} ; \mathbf{e}_i \rangle_\mathbf{M}$ on a :

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{bmatrix} = \mathbf{XMa} = \mathbf{Xu} = \sum_{j=1}^{p} \mathbf{x}^j u_j$$

en posant $\mathbf{u} = \mathbf{Ma}$.

A la variable **c** sont donc associés trois êtres mathématiques :

- un axe Δ de F de vecteur unitaire **a** ;
- un vecteur **c** de E espace des variables ;
- une forme linéaire **u** appelée facteur.

L'ensemble des variables **c** que l'on peut engendrer par combinaison linéaire des vecteurs-colonnes de **X** forme un sous-espace vectoriel de E de dimension égale (ou inférieure) à p.

Remarquons que si **a** appartient à l'espace des individus F, **u** appartient à son dual F^*, et que si **a** est **M**-normé à 1, **u** est $\mathbf{M}^{-1}$ normé à 1 :

$$\mathbf{a'Ma} = \mathbf{u'M}^{-1}\mathbf{u} \qquad \text{car } \mathbf{u} = \mathbf{Ma} \Rightarrow \mathbf{a} = \mathbf{M}^{-1}\mathbf{u}$$

F^* est donc muni de la métrique $\mathbf{M}^{-1}$.

(Lorsque $\mathbf{M} = \mathbf{I}$ ces distinctions disparaissent et on peut identifier totalement axes et facteurs).

La variance de **c** vaut alors :

$$\boxed{V(\mathbf{c}) = s_\mathbf{c}^2 = \mathbf{u'Vu}}$$

En effet : $\mathbf{c'Dc} = (\mathbf{Xu})'\mathbf{D}(\mathbf{Xu}) = \mathbf{u'X'DXu}$

7.2 L'ANALYSE

7.2.1 Projection des individus sur un sous-espace

Le principe de la méthode est d'obtenir une représentation approchée du nuage des n individus dans un sous-espace de dimension faible. Ceci s'effectue par projection ainsi que l'illustre la figure 7.2.

Le choix de l'espace de projection s'effectue selon le critère suivant qui revient à déformer le moins possible les distances en projection : le sous-espace de dimension k recherché est tel que la moyenne des carrés des distances entre projections soit la plus grande possible. (En effet, en projection les distances ne peuvent que diminuer). En d'autres termes il faut que l'inertie du nuage projeté sur le sous-espace F_k soit maximale.

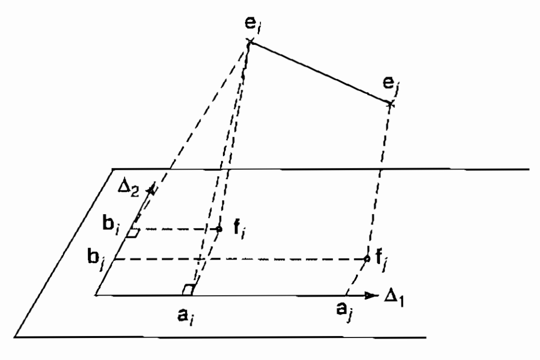


FIGURE 7.2

Soit **P** l'opérateur de projection **M**-orthogonale sur F_k : **P** est tel que $\mathbf{P}^2 = \mathbf{P}$ et $\mathbf{P'M} = \mathbf{MP}$.

Le nuage projeté est alors associé au tableau de données **XP'**, car chaque individu $\mathbf{e}_i$ (ou ligne de **X**) se projette sur F_k selon un vecteur colonne $\mathbf{Pe}_i$ ou un vecteur ligne $\mathbf{e}_i\mathbf{P}'$ (fig. 7.3).

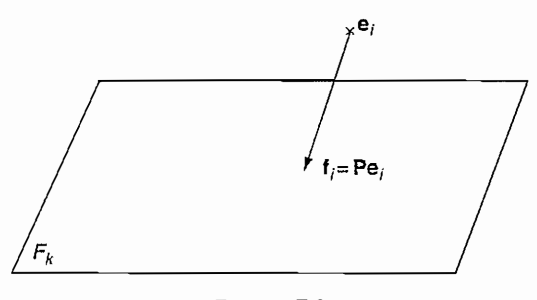


FIGURE 7.3

La matrice de variance du tableau **XP'** est pour des variables centrées :

$$(\mathbf{XP}')'\mathbf{D}(\mathbf{XP}') = \mathbf{PVP}'$$

L'inertie du nuage projeté vaut donc : Trace (**PVP'M**).

Par des opérations élémentaires on en déduit :

$$\begin{aligned}
\text{Trace}(\mathbf{PVP'M}) &= \text{Trace}(\mathbf{PVMP}) && \text{car } \mathbf{P'M} = \mathbf{MP} \\
&= \text{Trace}(\mathbf{VMP}^2) && \text{car Trace } \mathbf{AB} = \text{Trace } \mathbf{BA} \\
&= \text{Trace}(\mathbf{VMP}) && \text{car } \mathbf{P} \text{ est idempotent}
\end{aligned}$$

Le problème est donc de trouver **P**, projecteur **M**-orthogonal de rang k maximisant Trace (**VMP**) ce qui déterminera donc F_k.

Si F et G sont deux sous-espaces orthogonaux alors :

$$I_{F \oplus G} = I_F + I_G$$

Il suffit de remarquer que le projecteur associé à la somme directe de deux sous-espaces **M**-orthogonaux est la somme des projecteurs associés à chacun des espaces.

De ce résultat on déduit le théorème fondamental suivant :

THÉORÈME

*Soit F_k un sous-espace portant l'inertie maximale, alors le sous-espace de dimension $k+1$ portant l'inertie maximale est la somme directe de F_k et du sous-espace de dimension 1 **M**-orthogonal à F_k portant l'inertie maximale : Les solutions sont « emboîtées ».*

Démonstration : Soit E_{k+1} un sous-espace de dimension $k+1$.
Comme $\dim E_{k+1} = k+1$ et $\dim F_k^{\perp} = n-k$, on a :

$$\dim (E_{k+1} \cap F_k^{\perp}) \geq 1$$

car : $$\dim E_{k+1} + \dim F_k^{\perp} = n+1 > n$$

Soit $\mathbf{b}$ un vecteur appartenant à $E_{k+1} \cap F_k^{\perp}$.

Posons $E_{k+1} = \mathbf{b} \oplus G$ où G est le supplémentaire **M**-orthogonal de $\mathbf{b}$ dans E_{k+1}. G est donc de dimension k et $F = F_k \oplus \mathbf{b}$.

On a : $$I_{k+1} = I_{\mathbf{b}} + I_G$$
$$I_F = I_{F_k} + I_{\mathbf{b}}$$

Comme F_k était le sous-espace de dimension k portant l'inertie maximale $I_G \leq I_{Fk}$, donc $I_{k+1} \leq I_{\mathbf{b}} + I_{Fk}$, c'est-à-dire $I_{k+1} \leq I_F$ et ceci quel que soit E_{k+1}.

Le maximum de l'inertie est donc réalisé pour l'espace $F = \mathbf{b} \oplus F_k$ et $\mathbf{b}$ doit être tel que $I_{\mathbf{b}}$ soit maximal.

Pour obtenir F_k on pourra donc procéder de proche en proche en cherchant d'abord le sous-espace de dimension 1 d'inertie maximale, puis le sous-espace de dimension 1 **M**-orthogonal au précédent d'inertie maximale, etc.

7.2.2 Éléments principaux

7.2.2.1 *Axes principaux*

Nous devons chercher la droite de $\mathbb{R}^p$ passant par $\mathbf{g}$ maximisant l'inertie du nuage projeté sur cette droite.

Soit $\mathbf{a}$ un vecteur porté par cette droite ; le projecteur M-orthogonal sur la droite est alors :

$$\mathbf{P} = \mathbf{a}(\mathbf{a}'\mathbf{M}\mathbf{a})^{-1}\mathbf{a}'\mathbf{M}$$

L'inertie du nuage projeté sur cette droite vaut, d'après ce qui précède :

$$\text{Trace } \mathbf{VMP} = \text{Trace } \mathbf{VMa}(\mathbf{a}'\mathbf{Ma})^{-1}\mathbf{a}'\mathbf{M}$$
$$= \frac{1}{\mathbf{a}'\mathbf{Ma}} \text{Trace } \mathbf{VMaa}'\mathbf{M} = \frac{\text{Trace } \mathbf{a}'\mathbf{MVMa}}{\mathbf{a}'\mathbf{Ma}} = \frac{\mathbf{a}'\mathbf{MVMa}}{\mathbf{a}'\mathbf{Ma}}$$

puisque $\mathbf{a}'\mathbf{MVMa}$ est un scalaire.

La matrice **MVM** est appelée matrice d'inertie du nuage ; elle définit la forme quadratique d'inertie qui, à tout vecteur **a** de **M**-norme 1, associe l'inertie projetée sur l'axe défini par **a**. La matrice d'inertie ne se confond avec la matrice de variance-covariance que si $\mathbf{M} = \mathbf{I}$.

Pour obtenir le maximum de $\dfrac{\mathbf{a}'\mathbf{MVMa}}{\mathbf{a}'\mathbf{Ma}}$ il suffit d'annuler la dérivée de cette expression par rapport à **a** :

$$\frac{d}{d\mathbf{a}}\left(\frac{\mathbf{a}'\mathbf{MVMa}}{\mathbf{a}'\mathbf{Ma}}\right) = \frac{(\mathbf{a}'\mathbf{Ma})2\mathbf{MVMa} - (\mathbf{a}'\mathbf{MVMa})2\mathbf{Ma}}{(\mathbf{a}'\mathbf{Ma})^2}$$

d'où :

$$\mathbf{MVMa} = \left(\frac{\mathbf{a}'\mathbf{MVMa}}{\mathbf{a}'\mathbf{Ma}}\right)\mathbf{Ma}$$

soit :

$$\boxed{\mathbf{VMa} = \lambda\mathbf{a}}$$

car **M** est régulière ; donc **a** est vecteur propre de **VM**. S'il en est ainsi, le critère $\mathbf{a}'\mathbf{MVMa}$ vaut $\lambda\mathbf{a}'\mathbf{Ma} = \lambda$. Il faut donc que λ soit la plus grande valeur propre de **VM**.

La matrice **VM** étant **M**-symétrique possède des vecteurs propres **M**-orthogonaux deux à deux.

D'où le résultat suivant :

THÉORÈME

Le sous-espace F_k de dimension k est engendré par les k vecteurs propres de **VM** *associés aux k plus grandes valeurs propres.*

On appelle axes principaux d'inertie les vecteurs propres de **VM**, **M**-normés à 1. Ils sont au nombre de p.

Un calcul élémentaire montre que les axes principaux sont aussi $\mathbf{V}^{-1}$ orthogonaux ; on montre réciproquement que les axes principaux sont le seul système de vecteurs à la fois **M** et $\mathbf{V}^{-1}$-orthogonaux.

7.2.2.2 Facteurs principaux

A l'axe **a** est associée la forme linéaire **u** coordonnée **M**-orthogonale sur l'axe défini par **a** (fig. 7.4).

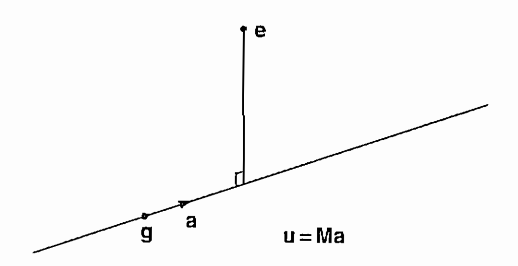


FIGURE 7.4

u est un élément de $(\mathbb{R}^p)^*$ (dual de l'espace des individus) qui définit une combinaison linéaire des variables descriptives $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$.

A l'axe principal **a** **M**-normé à 1 est associé le facteur principal $\mathbf{u} = \mathbf{Ma}$.

Puisque **a** était vecteur-propre de **VM** :

$$\mathbf{VMa} = \lambda\mathbf{a} \Rightarrow \mathbf{MVMa} = \lambda\mathbf{Ma}$$

soit :

$$\boxed{\mathbf{MVu} = \lambda\mathbf{u}}$$

Les facteurs principaux sont les vecteurs propres $\mathbf{M}^{-1}$-normés de **MV**. En effet, on a vu que si $\mathbb{R}^p$ est muni de la métrique **M**, son dual doit être muni de la métrique $\mathbf{M}^{-1}$. Donc $\mathbf{u}'\mathbf{M}^{-1}\mathbf{u} = 1$.

Les facteurs principaux sont $\mathbf{M}^{-1}$ et **V**-orthogonaux.

7.2.2.3 Composantes principales

Ce sont les variables $\mathbf{c}_i$ (éléments de $\mathbb{R}^n$) définies par les facteurs principaux :

$$\mathbf{c}_i = \mathbf{Xu}_i$$

$\mathbf{c}_i$ est le vecteur renfermant les coordonnées des projections **M**-orthogonales des individus sur l'axe défini par $\mathbf{a}_i$ avec $\mathbf{a}_i$ unitaire.

La variance d'une composante principale est égale à la valeur propre λ :

$$\boxed{V(\mathbf{c}_i) = \lambda_i}$$

En effet $V(\mathbf{c}) = \mathbf{c}'\mathbf{Dc} = \mathbf{u}'\mathbf{X}'\mathbf{DXu} = \mathbf{u}'\mathbf{Vu}$ or :

$$\mathbf{Vu} = \lambda\mathbf{M}^{-1}\mathbf{u}$$

donc :

$$V(\mathbf{c}) = \lambda\mathbf{u}'\mathbf{M}^{-1}\mathbf{u} = \lambda$$

Les c_i sont les combinaisons linéaires de $x_1, x_2, \ldots, x_p$ de variance maximale sous la contrainte $u_i'M^{-1}u_i = 1$.

Les composantes principales sont elles-mêmes vecteurs propres d'une matrice de taille n :

En effet $\mathbf{MVu} = \lambda\mathbf{u}$ s'écrit $\mathbf{MX'DXu} = \lambda\mathbf{u}$; en multipliant à gauche par $\mathbf{X}$ et en remplaçant $\mathbf{Xu}$ par $\mathbf{c}$ on a :

$$\boxed{\mathbf{XMX'Dc} = \lambda\mathbf{c}}$$

La matrice $\mathbf{XMX'}$ notée $\mathbf{W}$ est la matrice dont le terme général w_{ij} est le produit scalaire $\langle \mathbf{e}_i ; \mathbf{e}_j \rangle = \mathbf{e}_i'\mathbf{M}\mathbf{e}_j$.

D'où pour résumer :

Facteurs principaux $\mathbf{u}$	$\mathbf{MVu} = \lambda\mathbf{u}$	$\mathbf{M}^{-1}$-orthonormés
Axes principaux $\mathbf{a}$	$\mathbf{VMa} = \lambda\mathbf{a}$	$\mathbf{M}$-orthonormes
Composantes principales $\mathbf{c}$	$\mathbf{XMX'Dc} = \lambda\mathbf{c}$	$\mathbf{D}$-orthogonales
$\mathbf{c} = \mathbf{Xu}$	$\mathbf{u} = \mathbf{Ma}$	

En pratique on calcule les $\mathbf{u}$ par diagonalisation de $\mathbf{MV}$, puis on obtient les $\mathbf{c} = \mathbf{Xu}$, les axes principaux $\mathbf{a}$ n'ayant pas d'intérêt pratique.

7.2.2.4 Formules de reconstitution

Comme $\mathbf{Xu}_j = \mathbf{c}_j$ en post-multipliant les deux membres par $\mathbf{u}_j'\mathbf{M}^{-1}$ et en sommant sur j il vient :

$$\mathbf{X}\sum_j \mathbf{u}_j\mathbf{u}_j'\mathbf{M}^{-1} = \sum_j \mathbf{c}_j\mathbf{a}_j'\mathbf{M}^{-1}$$

Or $\sum_{j=1}^{p} \mathbf{u}_j\mathbf{u}_j'\mathbf{M}^{-1} = \mathbf{I}$ car les $\mathbf{u}_j$ sont $\mathbf{M}^{-1}$ orthonormés, il suffit de vérifier que :

$$\left(\sum_{j=1}^{p} \mathbf{u}_j\mathbf{u}_j\mathbf{M}^{-1}\right)\mathbf{u}_i = \mathbf{u}_i \quad \text{car} \quad \mathbf{u}_j'\mathbf{M}^{-1}\mathbf{u}_i = \delta_{ij}$$

donc :

$$\boxed{\mathbf{X} = \sum_{j=1}^{p} \mathbf{c}_j\mathbf{u}_j'\mathbf{M}^{-1}}$$

On peut ainsi reconstituer le tableau de données (centré) au moyen des composantes principales et facteurs principaux. On a également :

$$\boxed{\begin{aligned} \mathbf{MV} &= \sum_{j=1}^{p} \lambda_j\mathbf{u}_j\mathbf{u}_j'\mathbf{M}^{-1} \\ \mathbf{VM} &= \sum_{j=1}^{p} \lambda_j\mathbf{a}_j\mathbf{a}_j'\mathbf{M} \end{aligned}}$$

Lorsque $\mathbf{M} = \mathbf{I}, \mathbf{X} = \sum_{j=1}^{p} \mathbf{c}_j \mathbf{u}'_j = \sum_{j=1}^{p} \sqrt{\lambda_j} \mathbf{z}_j \mathbf{v}'_j$ où les $\mathbf{z}_j$ sont les vecteurs propres de norme 1 de $\mathbf{XX}'$ et les $\mathbf{v}_j$ les vecteurs propres de $\mathbf{X}'\mathbf{X}$ de norme 1 (décomposition en valeurs singulières).

Si l'on se contente de la somme des k premiers termes on obtient alors la meilleure approximation de $\mathbf{X}$ par une matrice de rang k au sens des moindres carrés (théorème d'Eckart-Young).

7.2.3 Cas usuel. La métrique $\mathbf{D}_{1/s^2}$ ou l'ACP sur données centrées-réduites

Le choix de la métrique $\mathbf{M}$ est toujours délicat : seul l'utilisateur peut définir correctement la notion de distance entre individus.

Prendre $\mathbf{M} = \mathbf{I}$ revient à travailler sur la matrice $\mathbf{V}$ des variances-covariances, il n'y a pas alors de distinction entre axes principaux et facteurs principaux. Cependant, les résultats obtenus ne sont pas invariants si l'on change linéairement l'unité de mesure des variables. Les covariances sont multipliées par un facteur k, la variance par un facteur k^2 si l'on choisit une unité de mesure k fois plus petite pour une variable.

Le choix de $\mathbf{M} = \mathbf{D}_{1/s^2}$ est le plus communément fait, et a pour conséquence de rendre les distances entre individus invariantes par transformation linéaire séparée de chaque variable et de s'affranchir des unités de mesure ce qui est particulièrement intéressant lorsque les variables sont hétérogènes.

On sait que l'usage de cette métrique est équivalent à la réduction des variables (division par l'écart-type).

En pratique on travaillera donc sur le tableau centré-réduit $\mathbf{Z}$ associé à $\mathbf{X}$ et on utilisera la métrique $\mathbf{M} = \mathbf{I}$.

Comme la matrice de variance-covariance des données centrées et réduites est la matrice de corrélation $\mathbf{R}$, les facteurs principaux seront donc les vecteurs propres successifs de $\mathbf{R}$ rangés selon l'ordre décroissant des valeurs propres. $\mathbf{Ru} = \lambda\mathbf{u}$ avec $\|\mathbf{u}\|^2 = 1$.

La première composante principale $\mathbf{c}$ (et les autres sous la contrainte d'orthogonalité) est la combinaison linéaire des variables centrées et réduites ayant une variance maximale $\mathbf{c} = \mathbf{Zu}$.

On a de plus la propriété suivante lorsqu'on travaille sur données centrées et réduites :

PROPRIÉTÉ

$\mathbf{c}$ est la variable la plus liée aux x^j au sens de la somme des carrés des corrélations :

$$\sum_{j=1}^{p} r^2(\mathbf{c} ; \mathbf{x}^j) \text{ est maximal}$$

Cette propriété permet de généraliser l'ACP à d'autres méthodes et d'autres type de variables en remplaçant le coefficient de corrélation par un indice adapté (principe d'association maximale, voir plus loin).

Démonstration : Supposons les variables centrées :

$r^2(\mathbf{c} ; \mathbf{x}^j) = r^2(\mathbf{c} ; \mathbf{z}^j)$ où $\mathbf{z}^j = \dfrac{\mathbf{x}^j}{s_j}$ est la variable centrée-réduite associée à $\mathbf{x}^j$:

$$r^2(\mathbf{c} ; \mathbf{z}^j) = \frac{[\text{cov}(\mathbf{c} ; \mathbf{z}^j)]^2}{V(\mathbf{c})V(\mathbf{z}^j)} = \frac{[\mathbf{c}'\mathbf{D}\mathbf{z}^j]^2}{V(\mathbf{c})}$$

$$\sum_{j=1}^{p} r^2(\mathbf{c} ; \mathbf{z}^j) = \frac{1}{V(\mathbf{c})}\sum_{j=1}^{p}(\mathbf{c}'\mathbf{D}\mathbf{z}^j)(\mathbf{z}^{j\prime}\mathbf{D}\mathbf{c}) = \frac{1}{V(\mathbf{c})}\mathbf{c}'\mathbf{D}\left(\sum_{j=1}^{p}\mathbf{z}^j\mathbf{z}^{j\prime}\right)\mathbf{D}\mathbf{c}$$

or :

$$\sum_{j=1}^{p}\mathbf{z}^j(\mathbf{z}^j)' = \mathbf{Z}\mathbf{Z}'$$

donc :

$$\sum_{j=1}^{p} r^2(\mathbf{c} ; \mathbf{x}^j) = \frac{\mathbf{c}'\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c}}{\mathbf{c}'\mathbf{D}\mathbf{c}}$$

le maximum de ce quotient est donc atteint pour $\mathbf{c}$ vecteur propre de $\mathbf{Z}\mathbf{Z}'\mathbf{D}$ associé à sa plus grande valeur propre :

$$\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c} = \lambda\mathbf{c}$$

on en déduit que $\mathbf{c}$ est combinaison linéaire des $\mathbf{z}^j$ donc que $\mathbf{c} = \mathbf{Z}\mathbf{u}$;

$$\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{Z}\mathbf{u} = \lambda\mathbf{Z}\mathbf{u}$$

Comme $\mathbf{Z}'\mathbf{D}\mathbf{Z} = \mathbf{R}$, il vient $\mathbf{Z}\mathbf{R}\mathbf{u} = \lambda\mathbf{Z}\mathbf{u}$ et si $\mathbf{Z}$ est de rang p : $\mathbf{R}\mathbf{u} = \lambda\mathbf{u}$. ■

Pour résumer : l'ACP revient à remplacer les variables $x^1, x^2, \ldots, x^p$ qui sont corrélées, par de nouvelles variables, les composantes principales $c^1, c^2, \ldots$ combinaisons linéaires des x^j non corrélées entre elles, de variance maximale et les plus liées en un certain sens aux x^j : l'ACP est une méthode factorielle linéaire.

7.3 INTERPRÉTATION DES RÉSULTATS

L'ACP construit de nouvelles variables, artificielles et fournit des représentations graphiques permettant de visualiser les relations entre variables ainsi que l'existence éventuelle de groupes d'individus et de groupes de variables.

L'interprétation des résultats est une phase délicate qui doit se faire en respectant une démarche dont les éléments sont les suivants.

7.3.1 Qualité des représentations sur les plans principaux

Le but de l'ACP étant d'obtenir une représentation des individus dans un espace de dimension plus faible que p, la question se pose d'apprécier la perte d'information subie et de savoir combien de facteurs retenir.

7.3.1.1 Le pourcentage d'inertie

Le critère habituellement utilisé est celui du pourcentage d'inertie totale expliquée.

On mesure la qualité de F_k par :

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{I_g} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \cdots + \lambda_p}$$

Si par exemple $\dfrac{\lambda_1 + \lambda_2}{I_g} = 0.9$ on conçoit clairement que le nuage de points est presque aplati sur un sous-espace à deux dimensions et qu'une représentation du nuage dans le plan des deux premiers axes principaux sera très satisfaisante.

L'appréciation du pourcentage d'inertie doit faire intervenir le nombre de variables initiales : un % de 10 % n'a pas le même intérêt sur un tableau de 20 variables et sur un tableau de 100 variables.

7.3.1.2 Mesures locales

Le pourcentage d'inertie expliquée est un critère global qui doit être complété par d'autres considérations.

Supposons que le plan F_2 des deux premiers axes porte une inertie totale importante ($\lambda_1 + \lambda_2$ élevé) et que en projection sur ce plan deux individus soient très proches : la figure 7.5 montre que cette proximité peut être illusoire si les deux individus se trouvent éloignés dans $F_2^{\perp}$.

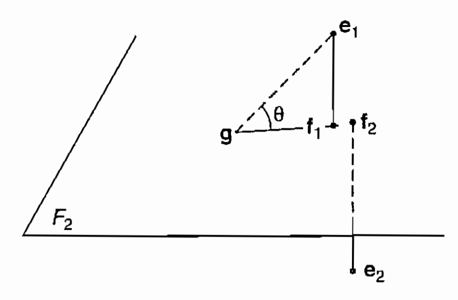


FIGURE 7.5

Il faut en fait envisager pour chaque individu $\mathbf{e}_i$ la qualité de sa représentation. Celle-ci est souvent définie par le cosinus de l'angle entre le plan principal et le vecteur $\mathbf{e}_j$. Si ce cosinus est grand, $\mathbf{e}_i$ est voisin du plan, on pourra alors examiner la position de sa projection sur le plan par rapport à d'autres points ; si ce cosinus est faible on se gardera de toute conclusion.

N.B. : Cette mesure du cosinus est d'autant meilleure que $\mathbf{e}_i$ est éloigné de $\mathbf{g}$; si $\mathbf{e}_i$ est proche de $\mathbf{g}$, la valeur du cosinus peut ne pas être significative.

Bien que moins utilisée, une mesure liée à la distance entre $\mathbf{e}_i$ et F_k semble préférable : en particulier la quantité :

$$\frac{d(\mathbf{e}_i\,;\mathbf{f}_i)}{\sqrt{I_g - \lambda_1 - \lambda_2 - \cdots - \lambda_k}}\ (\text{signe de } c_i^{k+1})$$

qui compare la distance entre $\mathbf{e}_i$ et F_k à la moyenne des carrés des distances de tous les individus à F_k présente un intérêt statistique certain (on peut la comparer à une variable de Laplace-Gauss centrée-réduite).

7.3.1.3 A propos de la représentation simultanée des individus et des variables en ACP

Certains logiciels prévoient la possibilité de superposer la représentation des individus (plan principal) et celle des variables (cercle des corrélations) avec éventuellement des échelles différentes.

Il convient d'être très prudent : en effet individus et variables sont des éléments d'espaces différents : si une variable définit une direction de l'espace des individus elle ne peut être résumée à un point et on ne peut interpréter une proximité entre points-variables et points-individus.

Les deux représentations individus et variables se complètent mais ne peuvent être superposées, sauf en utilisant la technique particulière du "biplot" (voir Gower et Hand, 1996).

7.3.2 Choix de la dimension

Le principal intérêt de l'ACP consistant à réduire la dimension de l'espace des individus le choix du nombre d'axes à retenir est un point essentiel qui n'a pas de solution rigoureuse. Remarquons tout d'abord que la réduction de dimension n'est possible que s'il y a redondance entre les variables $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^p$: si celles-ci sont indépendantes, ce qui est un résultat fort intéressant en soi, l'ACP sera inefficace à réduire la dimension.

7.3.2.1 Critères théoriques

Ceux-ci consistent à déterminer si les valeurs propres sont significativement différentes entre elles à partir d'un certain rang : si la réponse est négative on conserve les premières valeurs propres. On fait pour cela l'hypothèse que les n individus proviennent d'un tirage aléatoire dans une population gaussienne où $\lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p$. Si cela est vrai la moyenne arithmétique a des $k - p$ dernières valeurs propres et leur moyenne géométrique g doivent être peu différentes ; on admet que :

$$c = \left(n - \frac{2p + 11}{6}\right)(p - k)\ln\left(\frac{a}{g}\right)$$

suit alors une loi du χ^2 de degré de liberté $\frac{(p - k + 2)(p - k - 1)}{2}$; on rejettera l'hypothèse d'égalité des $k - p$ valeurs propres si c est trop grand.

On peut également construire des intervalles de confiance pour les différentes valeurs propres en utilisant les formules de T. W. Anderson si n est grand : si λ_i est la $i^{\text{ième}}$ valeur propre de l'ACP, l'intervalle de confiance à 95 % est donné par :

$$\left[\lambda_i \exp\left(-1.96\sqrt{\frac{2}{n-1}}\right) ; \lambda_i \exp\left(1.96\sqrt{\frac{2}{n-1}}\right)\right] \qquad \text{(voir chapitre 13)}$$

Ces propriétés ne sont malheureusement utilisables que pour des matrices de variance dans le cas gaussien p-dimensionnel. Elles ne s'appliquent pas pour les matrices de corrélation ce qui est le cas le plus fréquent en pratique, et ne doivent être utilisées qu'à titre indicatif.

7.3.2.2 Critères empiriques

Ce sont en réalité les seuls applicables, le critère de Kaiser est le plus connu :

Lorsqu'on travaille sur données centrées réduites on retient les composantes principales correspondant à des valeurs propres supérieures à 1 : en effet les composantes principales $\mathbf{c}$ étant des combinaisons linéaires des $\mathbf{z}^j$ de variance maximale $V(\mathbf{c}) = \lambda$, seules les composantes de variance supérieure à celle des variables initiales présentent un intérêt.

Cependant le seuil de 1 ne peut être considéré comme absolu : 1.1 est-il significativement supérieur à 1 ?

Dans un travail récent (Karlis, Saporta, Spinakis, 2003) nous avons montré l'intérêt du critère suivant, inspiré par une approche de type « carte de contrôle » où on considère comme intéressantes les valeurs propres qui dépassent leur moyenne (qui vaut ici 1) de plus de deux écart-types.

Comme :
$$\sum_{i=1}^{p} \lambda_i^2 = p + \sum_{i \neq j}\sum r_{ij}^2$$

et que l'espérance du carré du coefficient de corrélation entre deux variables indépendantes vaut $1/(n-1)$, on trouve que :

$$E\left(\sum_{i=1}^{p} \lambda_i^2\right) = p + \frac{p(p-1)}{n-1}$$

la dispersion espérée des valeurs propres vaut alors :

$$E\left(\frac{1}{p}\sum_{i=1}^{p}(\lambda_i - 1)^2\right) = \frac{p-1}{n-1}$$

Nous proposons donc de ne retenir que les valeurs propres telles que :

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}.$$

On recommande également de détecter sur le diagramme de décroissance des valeurs propres l'existence d'un coude (voir figure 7.8) séparant les valeurs propres utiles de celles qui sont peu différentes entre elles et n'apportent pas d'information. Il faut noter ici que les critères formels basés sur les différences successives entre valeurs propres sont en général moins performants que l'inspection visuelle : nous ne les donnerons donc pas.

Enfin il faut rappeler avec force que les critères du type « extraire au moins x % » de l'inertie souvent prônés par des praticiens, sont dénués de fondement et doivent être bannis, car on ne peut donner de seuil universel sans tenir compte de la taille du tableau, et de la force des corrélations entre variables.

Aucun des critères présentés n'est absolu : l'interprétation des résultats d'une analyse relève aussi du métier du statisticien.

7.3.3 Interprétation « interne »

7.3.3.1 Corrélations « variables – facteurs »

La méthode la plus naturelle pour donner une signification à une composante principale $\mathbf{c}$ est de la relier aux variables initiales $\mathbf{x}^j$ en calculant les coefficients de corrélation linéaire $r(\mathbf{c} ; \mathbf{x}^j)$ et en s'intéressant aux plus forts coefficients en valeur absolue.

Lorsque l'on choisit la métrique $\mathbf{D}_{1/s^2}$ ce qui revient à travailler sur données centrées-réduites et donc à chercher les valeurs propres et vecteurs propres de $\mathbf{R}$, le calcul de $r(\mathbf{c} ; \mathbf{x}^j)$ est particulièrement simple :

En effet :

$$r(\mathbf{c} ; \mathbf{x}^j) = r(\mathbf{c} ; \mathbf{z}^j) = \frac{\mathbf{c}'\mathbf{D}\mathbf{z}^j}{s_\mathbf{c}}$$

comme $V(\mathbf{c}) = \lambda$:

$$r(\mathbf{c} ; \mathbf{x}^j) = \frac{\mathbf{c}'\mathbf{D}\mathbf{z}^j}{\sqrt{\lambda}}$$

or $\mathbf{c} = \mathbf{Z}\mathbf{u}$ où $\mathbf{u}$, facteur principal associé à $\mathbf{c}$, est vecteur propre de $\mathbf{R}$ associé à la valeur propre λ :

$$r(\mathbf{c} ; \mathbf{x}^j) = \mathbf{u}'\mathbf{Z}'\mathbf{D}\mathbf{z}^j = \frac{(\mathbf{z}^j)'\mathbf{D}\mathbf{Z}\mathbf{u}}{\sqrt{\lambda}}$$

$(\mathbf{z}^j)'\mathbf{D}\mathbf{Z}$ est la $j^{\text{ème}}$ ligne de $\mathbf{Z}'\mathbf{D}\mathbf{Z} = \mathbf{R}$, donc $(\mathbf{z}^j)'\mathbf{D}\mathbf{Z}\mathbf{u}$ est la $j^{\text{ème}}$ composante de $\mathbf{R}\mathbf{u}$. Comme $\mathbf{R}\mathbf{u} = \lambda\mathbf{u}$, il vient :

$$\boxed{r(\mathbf{c} ; \mathbf{x}^j) = \sqrt{\lambda}\, u_j}$$

Ces calculs s'effectuent pour chaque composante principale. Pour un couple de composantes principales $\mathbf{c}^1$ et $\mathbf{c}^2$ par exemple on synthétise usuellement les corrélations sur une figure appelée « **cercle des corrélations** » où chaque variable $\mathbf{x}^j$ est repérée par un point d'abscisse $r(\mathbf{c}^1 ; \mathbf{x}^j)$ et d'ordonnée $r(\mathbf{c}^2 ; \mathbf{x}^j)$.

Ainsi la figure 7.6 montre une première composante principale très corrélée positivement avec les variables 1, 2 et 3, anticorrélée avec les variables 4 et 5 et non corrélée avec 6, 7 et 8.

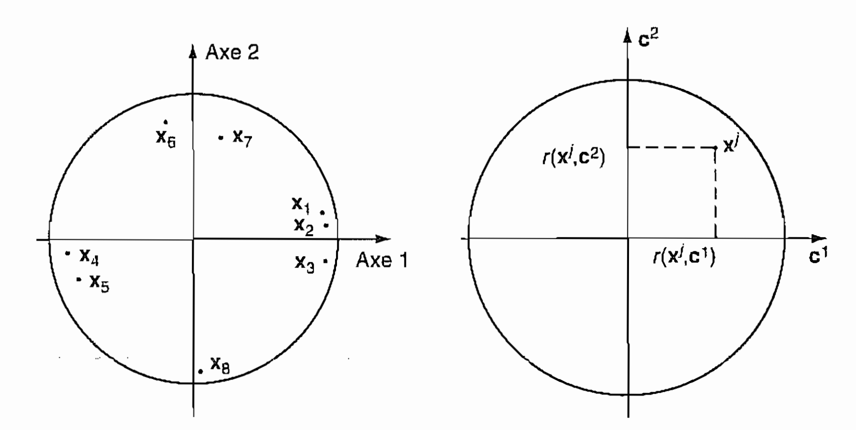


FIGURE 7.6

Par contre la deuxième composante principale oppose la variable n° 8 aux variables 6 et 7.

On se gardera d'interpréter des proximités entre points variables, si ceux-ci ne sont pas proches de la circonférence.

Dans le cas de la métrique $\mathbf{D}_{1/s^2}$ c'est-à-dire, rappelons-le, de l'ACP sur données centrées réduites, le cercle des corrélations n'est pas seulement une représentation symbolique commode : c'est la projection de l'ensemble des variables centrées-réduites sur le sous-espace engendré par $\mathbf{c}^1$ et $\mathbf{c}^2$. En effet les $\mathbf{z}^j$ étant de variance un, sont situées sur la surface de la sphère unité de l'espace des variables (isomorphe à $\mathbb{R}^n$) (fig. 7.7). Projetons les extrémités des vecteurs $\mathbf{z}^j$ sur le sous-espace de dimension 2 engendré par $\mathbf{c}^1$ et $\mathbf{c}^2$ (qui sont orthogonales) les projections tombent à l'intérieur du grand cercle intersection de la sphère avec le plan $\mathbf{c}^1$; $\mathbf{c}^2$. La projection se faisant avec la métrique $\mathbf{D}$ de l'espace des variables, $\mathbf{z}$ se projette sur l'axe engendré par $\mathbf{c}^1$ en un point d'abscisse cos $(\mathbf{z}^j ; \mathbf{c}^1)$ qui n'est autre que le coefficient de corrélation linéaire $r(\mathbf{x}^j ; \mathbf{c}^1)$.

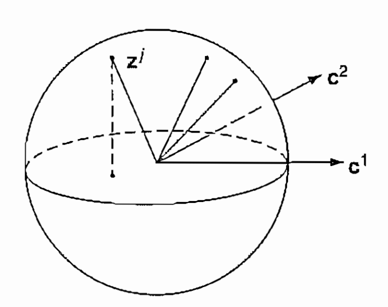


FIGURE 7.7

Le cercle de corrélation est donc, dans l'espace des variables, le pendant exact de la projection des individus sur le premier plan principal.

Comme $\lambda_k = \sum_{j=1}^{p} r^2(\mathbf{c}^k ; \mathbf{x}^j)$ on appelle parfois contribution de la variable j à l'axe k le rapport :

$$\frac{r^2(\mathbf{c}^k ; \mathbf{x}^j)}{\lambda_k} = (u_j^k)^2$$

mais cette quantité ne présente que peu d'intérêt en ACP et n'apporte rien de plus que le coefficient de corrélation.

7.3.3.2 La place et l'importance des individus

Dire que $\mathbf{c}^1$ est très corrélée avec une variable $\mathbf{x}^j$ signifie que les individus ayant une forte coordonnée positive sur l'axe 1 sont caractérisés par une valeur de $\mathbf{x}^j$ nettement supérieure à la moyenne (rappelons que l'origine des axes principaux représente le centre de gravité du nuage).

Inversement si les individus ne sont pas anonymes, ils aident à l'interprétation des axes principaux et des composantes principales : on recherchera par exemple les individus opposés le long d'un axe.

Il est très utile aussi de calculer pour chaque axe la contribution apportée par les divers individus à cet axe. Considérons la $k^{\text{ième}}$ composante $\mathbf{c}_k$; soit c_{ki} la valeur de cette composante pour le $i^{\text{ème}}$ individu. On a :

$$\sum_{i=1}^{n} p_i \, c_{ki}^2 = \lambda_k$$

La ***contribution*** de l'individu i à la composante $\mathbf{c}_k$ est définie par :

$$\frac{p_i \, c_{ki}^2}{\lambda_k}$$

La considération des contributions, quand elles ne sont pas excessives, aide à l'interprétation des axes.

Normalement, et ceci surtout pour les premières composantes, il n'est pas souhaitable qu'un individu ait une contribution excessive car cela serait un facteur d'instabilité, le fait de retirer cet individu modifiant profondément le résultat de l'analyse. Si ce cas se produisait il y aurait intérêt à effectuer l'analyse en éliminant cet individu puis en le mettant en élément supplémentaire, s'il ne s'agit pas d'une donnée erronée (erreur de saisie . . .) qui a été ainsi mise en évidence.

Cette remarque est surtout valable lorsque les individus constituent un échantillon et ne présentent donc pas d'intérêt en eux-mêmes.

Lorsque les poids des individus sont tous égaux à $1/n$ les contributions n'apportent pas plus d'information que les coordonnées.

Lorsque n est grand, il est souvent possible de considérer que les coordonnées sur une composante principale (du moins pour les premières composantes) qui ont pour moyenne 0 et pour

variance la valeur propre, sont distribuées selon une loi de Laplace-Gauss. Alors $\frac{c_{ik}^2}{\lambda_k}$ est distribué comme un χ^2 à un degré de liberté et la contribution $\frac{1}{n}\frac{c_{ik}^2}{\lambda_k}$ a une probabilité 0.05 de dépasser $3.84/n$. On pourra donc considérer qu'un individu a une contribution significative si elle dépasse 4 fois son poids.

7.3.3.3 Effet « taille »

Lorsque toutes les variables $\mathbf{x}_j$ sont corrélées positivement entre elles, la première composante principale définit un « facteur de taille ».

On sait qu'une matrice symétrique ayant tous ses termes positifs admet un premier vecteur propre dont toutes les composantes sont de même signe (théorème de Frobenius) : si l'on les choisit positives la première composante principale est alors corrélée positivement avec toutes les variables et les individus sont rangés sur l'axe 1 par valeurs croissantes de l'ensemble des variables (en moyenne). Si de plus les corrélations entre variables sont toutes de même ordre la première composante principale est proportionnelle à la moyenne des variables initiales :

$$\frac{1}{p}\sum_{j=1}^{p}\mathbf{x}^j$$

La deuxième composante principale différencie alors des individus de « taille » semblable : on l'appelle facteur de « forme ».

7.3.4 Interprétation externe : variables et individus supplémentaires, valeur-test

Les interprétations fondées sur les remarques précédentes présentent le défaut d'être tautologiques : on explique les résultats à l'aide des données qui ont servi à les obtenir.

On risque de prendre pour une propriété des données ce qui pourrait n'être qu'un artefact dû à la méthode : il n'est pas étonnant par exemple de trouver de fortes corrélations entre la première composante principale $\mathbf{c}^1$ et certaines variables puisque $\mathbf{c}^1$ maximise :

$$\sum_{j=1}^{p} r^2(\mathbf{c}\,;\,\mathbf{x}^j)$$

On n'est donc pas sûr d'avoir découvert un phénomène significatif.

Par contre si l'on trouve une forte corrélation entre une composante principale et une variable qui n'a pas servi à l'analyse, le caractère probant de ce phénomène sera bien plus élevé. D'où la pratique courante de partager en deux groupes l'ensemble des variables : d'une part les variables « actives » qui servent à déterminer les axes principaux, d'autre part les variables « passives » ou supplémentaires que l'on relie *a posteriori* aux composantes principales.

On distinguera le cas des variables numériques supplémentaires de celui des variables qualitatives supplémentaires.

Les variables numériques supplémentaires peuvent être placées dans les cercles de corrélation : il suffit de calculer le coefficient de corrélation entre chaque variable supplémentaire

y et les composantes principales $\mathbf{c}^1$, $\mathbf{c}^2$ .... On peut alors utiliser les résultats du chapitre précédent pour détecter une corrélation significative.

Une variable qualitative supplémentaire correspond à la donnée d'une partition des n individus en k catégories : on peut faire apparaître par des symboles différents les individus de chaque catégorie sur les plans principaux. En général on se contente de représenter chaque catégorie par son centre de gravité : on peut alors mesurer au moyen du rapport de corrélation la liaison entre une variable qualitative supplémentaire et une composante principale et vérifier son caractère significatif au moyen du F de Fisher-Snedecor (voir chapitre 6).

Cependant la pratique la plus efficace consiste à calculer ce que l'on appelle la valeur-test associée à chaque modalité ou catégorie supplémentaire qui mesure sur chaque axe la différence entre la moyenne des individus concernés et la moyenne générale (nulle par construction puisque les composantes principales sont centrées). Plus précisément il s'agit de la différence divisée par l'écart-type correspondant au raisonnement suivant (voir chapitre 12 et 20) : si les n_i individus de la catégorie i étudiée avaient été tirés au hasard avec probabilités égales parmi les n de l'ensemble étudié, la moyenne de leurs coordonnées sur l'axe $n^\circ k$ serait une variable aléatoire d'espérance nulle et de variance $\frac{\lambda_k}{n_i}\frac{n-n_i}{n-1}$ car le tirage est sans remise. La valeur-test associée à la coordonnée a_{ik} du centre de gravité est alors :

$$\frac{a_{ik}}{\sqrt{\frac{\lambda_k}{n_i}\frac{n-n_i}{n-1}}}.$$

En se référant à la loi de Laplace-Gauss, ce qui se justifie si n_i est assez grand, on décidera qu'une modalité occupe une position significativement différente de la moyenne générale si en valeur absolue, la valeur-test dépasse 2 voire 3.

On peut également ne pas faire participer à l'analyse une partie des individus (on calcule les corrélations sans eux) ce qui permettra de vérifier sur cet échantillon-test des hypothèses formulées après une ACP sur les individus actifs. Il est d'ailleurs immédiat de positionner de nouveaux individus sur les axes principaux puisqu'il suffit de calculer des combinaisons linéaires de leurs caractéristiques.

7.4 EXEMPLE

Les données concernent les caractéristiques de 18 véhicules (anciens...) et figurent dans le tableau 17.1 page 428.

Pour l'analyse en composantes principales, les variables « finition » (qualitative) et « prix » ont été mises en éléments supplémentaires, il y a donc 6 variables actives.

7.4.1 Valeurs propres

Comme les variables sont exprimées avec des unités différentes, on effectue l'ACP sur données centrées réduites, ce qui conduit à chercher les valeurs et vecteurs propres de la matrice de corrélation **R** présentée en 6.1.2.3.

Les calculs ont été effectués avec le logiciel SPAD version 5.6

DIAGRAMME DES 6 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	4.4209	73.68	73.68	***
2	0.8561	14.27	87.95	***************
3	0.3731	6.22	94.17	*******
4	0.2139	3.57	97.73	****
5	0.0928	1.55	99.28	**
6	0.0433	0.72	100.00	*

L'application des critères de choix de dimension (§ 7.3.2) ne conduirait à retenir qu'une seule valeur propre, ce qui montre bien leurs limites. Nous conserverons deux dimensions représentant 88 % de l'inertie. Remarquons que les intervalles d'Anderson des valeurs propres suivantes sont tous en dessous de 1.

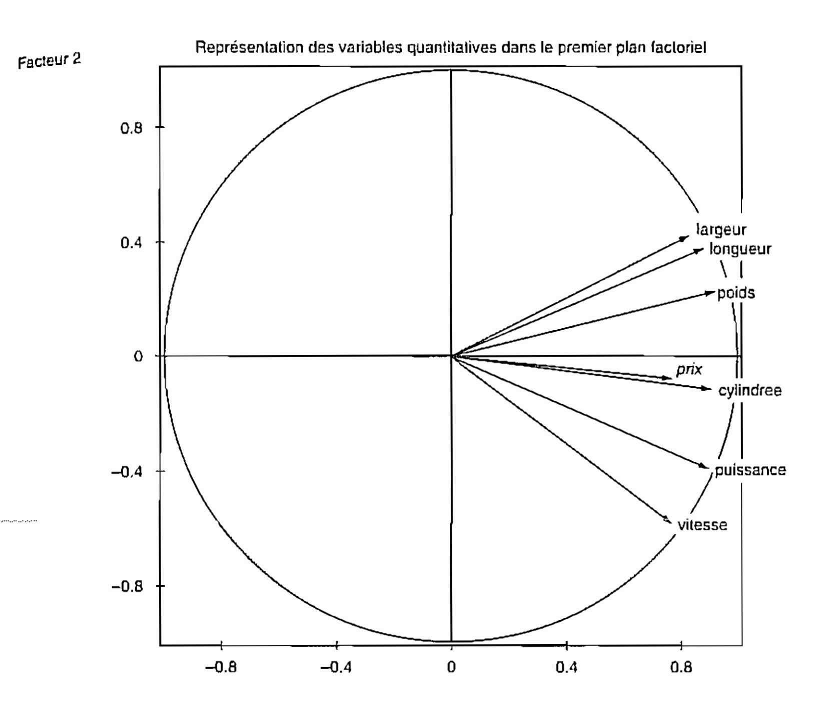
INTERVALLES A 0.95

NUMERO	BORNE INFERIEURE	VALEUR PROPRE	BORNE SUPERIEURE
1	1.4488	4.4209	7.3929
2	0.2806	0.8561	1.4316
3	0.1223	0.3731	0.6239
4	0.0701	0.2139	0.3577
5	0.0304	0.0928	0.1552

7.4.2 Interprétation des axes

Le tableau suivant ainsi que le cercle des corrélations montrent un effet « taille » sur le premier axe qui va donc classer les individus selon leur taille. Le deuxième axe s'interprète aisément comme opposant les véhicules sportifs aux autres.

VARIABLES	CORRELATIONS VARIABLE-FACTEUR				
IDEN - LIBELLE COURT	1	2	3	4	5
Cyli - cylindrée	0.89	−0.11	0.22	−0.37	−0.05
Puis - puissance	0.89	−0.38	0.11	0.17	0.09
Long - longueur	0.89	0.38	−0.04	0.13	−0.22
Larg - largeur	0.81	0.41	−0.37	−0.10	0.15
Poid - poids	0.91	0.22	0.30	0.14	0.09
Vite - vitesse	0.75	−0.57	−0.30	0.03	−0.06
Prix - prix	0.77	−0.09	0.13	0.23	0.16



La prise en compte des variables supplémentaires montre en outre que la première composante principale est liée à la qualité et au prix.

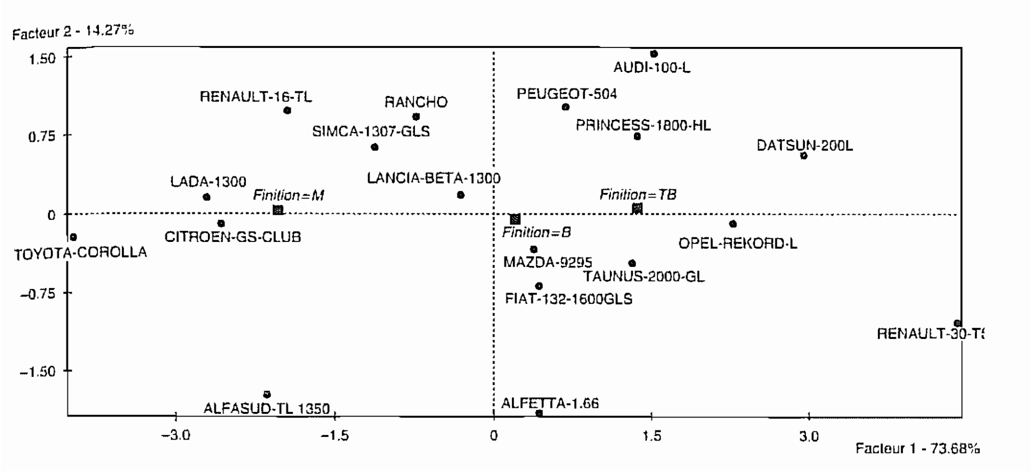
MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
finition													
F=B - Finition=B	7	7.00	0.4	−0.2	−0.6	−1.6	−0.8	0.24	−0.05	−0.11	−0.22	−0.08	0.12
F=M - Finition=M	5	5.00	−2.4	0.1	0.3	−0.3	−0.5	−2.00	0.02	0.07	−0.06	−0.06	4.02
F=TB - Finition=TB	6	6.00	1.9	0.1	0.4	1.9	1.3	1.39	0.03	0.07	0.30	0.14	2.06

7.4.3 Plan principal

Le tableau suivant fournit les composantes principales et les indices associés. Les individus les plus influents sur l'axe 1 sont RENAULT-30-TS et TOYOTA-COROLLA qui s'opposent par leur taille et sur l'axe 2 ALFASUD-TI-1350 et ALFETTA-1.66 , véhicules sportifs italiens.

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
ALFASUD-TI-1350	5.56	8.23	−2.14	−1.79	−0.57	−0.20	0.30	5.7	20.7	4.9	1.1	5.4	0.56	0.39	0.04	0.00	0.01
AUDI-100-L	5.56	6.67	1.56	1.53	−1.32	0.21	−0.15	3.1	15.1	25.8	1.2	1.3	0.37	0.35	0.26	0.01	0.00
SIMCA-1307-GLS	5.56	2.16	−1.12	0.67	−0.46	0.17	0.38	1.6	3.0	3.1	0.7	8.4	0.58	0.21	0.10	0.01	0.07
CITROEN-GS-CLUB	5.56	6.78	−2.57	−0.11	−0.15	0.02	−0.23	8.3	0.1	0.3	0.0	3.1	0.98	0.00	0.00	0.00	0.01
FIAT-132-1600GLS	5.56	1.17	0.43	−0.70	0.19	0.63	−0.26	0.2	3.1	0.6	10.2	4.2	0.16	0.41	0.03	0.34	0.06
LANCIA-BETA-1300	5.56	1.13	−0.30	0.20	−0.68	0.56	0.45	0.1	0.2	6.8	8.0	11.9	0.08	0.03	0.40	0.27	0.17
PEUGEOT-504	5.56	1.51	0.68	0.93	0.26	−0.20	−0.21	0.6	5.6	1.0	1.1	2.6	0.31	0.58	0.04	0.03	0.03
RENAULT-16-TL	5.56	5.64	−1.95	0.98	0.62	−0.63	−0.29	4.8	6.2	5.7	10.3	5.1	0.67	0.17	0.07	0.07	0.02
RENAULT-30-TS	5.56	21.79	4.41	−1.06	0.59	−0.85	0.37	24.4	7.3	5.2	18.6	8.4	0.89	0.05	0.02	0.03	0.01
TOYOTA-COROLLA	5.56	16.29	−3.99	−0.24	0.30	−0.27	−0.28	20.0	0.4	1.4	1.8	4.6	0.98	0.00	0.01	0.00	0.00
ALFETTA-1.66	5.56	4.46	0.14	−1.91	−0.02	0.76	−0.17	0.2	23.7	0.0	15.0	1.7	0.04	0.82	0.00	0.13	0.01
PRINCESS-1800-HL	5.56	1.95	1.02	0.84	−0.22	−0.30	0.18	1.3	4.6	0.7	2.4	2.0	0.53	0.36	0.02	0.05	0.02
DATSUN-200L	5.56	11.11	2.94	0.56	1.24	0.77	−0.05	10.9	2.0	23.0	15.5	0.2	0.78	0.03	0.14	0.05	0.00
TAUNUS-2000-GL	5.56	2.45	1.31	−0.49	−0.28	−0.58	0.07	2.2	1.5	1.2	8.8	0.3	0.70	0.10	0.03	0.14	0.00
RANCHO	5.56	1.96	-0.69	0.90	0.63	0.36	0.38	0.6	5.2	5.9	3.3	8.5	0.24	0.41	0.20	0.07	0.07
MAZDA-9295	5.56	0.68	0.39	−0.36	0.08	−0.10	−0.53	0.2	0.8	0.1	0.3	16.6	0.22	0.19	0.01	0.02	0.41
OPEL-REKORD-L	5.56	6.08	2.29	−0.10	−0.80	−0.24	−0.34	6.6	0.1	9.4	1.5	6.9	0.86	0.00	0.10	0.01	0.02
LADA-1300	5.56	7.92	−2.71	0.14	0.57	−0.10	0.38	9.2	0.1	4.9	0.2	8.7	0.93	0.00	0.04	0.00	0.02

Le plan principal donne la projection des 18 individus ainsi que les barycentres des modalités de la variable « Finition ».



7.5 ANALYSE FACTORIELLE SUR TABLEAUX DE DISTANCE ET DE DISSIMILARITÉS

Ces méthodes (***multidimensional scaling***) ont le même objectif que l'ACP : trouver une configuration de n individus dans un espace de faible dimension, mais les données de départ sont différentes ; ici on ne connaît que les $\frac{n(n-1)}{2}$ distances, ou dissimilarités entre individus, et non les variables les décrivant. Le cas où l'on dispose d'une véritable distance euclidienne entre individus n'est qu'une version de l'ACP, le cas de dissimilarités conduit à des techniques originales.

7.5.1 Analyse d'un tableau de distances euclidiennes

7.5.1.1 *La solution classique*

Soit Δ le tableau $n \times n$ des carrés des distances entre points :

$$d_{ij}^2 = d_{ji}^2 \qquad \text{et} \qquad d_{ii} = 0$$

Si d est euclidienne, chaque individu peut être représenté dans un espace de dimension p (pour l'instant inconnue) par un point $\mathbf{e}$ tel que :

$$d_{ij}^2 = (\mathbf{e}_i - \mathbf{e}_j)'(\mathbf{e}_i - \mathbf{e}_j)$$

On peut en effet toujours supposer $\mathbf{M} = \mathbf{I}$ sinon on s'y ramène par la transformation $\mathbf{T}$ telle que $\mathbf{M} = \mathbf{T'T}$. Si l'on place l'origine au centre de gravité, les produits scalaires $w_{ij} = \langle \mathrm{e}_i ; \mathrm{e}_j \rangle$ sont alors entièrement déterminés par les d_{ij}^2.

Supposons $p_i = 1/n \quad \forall i$ et posons $d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$ et $d_{..}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{i.}^2 = 2I$ où I est l'inertie du nuage.

On a alors la formule de Torgerson :

$$\boxed{w_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2)}$$

En effet :

$$d_{ij}^2 = \|\mathbf{e}_i\|^2 + \|\mathbf{e}_j\|^2 - 2w_{ij} \qquad \text{soit } w_{ij} = -\frac{1}{2}(-d_{ij}^2 + \|\mathbf{e}_i\|^2 + \|\mathbf{e}_j\|^2)$$

d'où :

$$d_{i.}^2 = \|\mathbf{e}_i\|^2 + \frac{1}{n}\sum_j \|\mathbf{e}_j\|^2 \qquad \text{car} \qquad \sum_j w_{ij} = \langle \mathbf{e}_i \,;\, \sum_j \mathbf{e}_j \rangle = 0$$

car l'origine est au centre de gravité.

On a donc $d_{i.}^2 = \|\mathbf{e}_i\|^2 + I$ et de même $d_{.j}^2 = \|\mathbf{e}_j\|^2 + I$ d'où la formule par substitution.

Matriciellement $\mathbf{W} = -\frac{1}{2}\mathbf{A}\Delta\mathbf{A}$ où $\mathbf{A}$ est l'opérateur de centrage $\mathbf{A} = \mathbf{I} - \frac{\mathbf{11}'}{n}$: il y a donc double centrage en ligne et en colonnes de Δ.

On sait que les vecteurs propres de $\mathbf{WD}$ $\left(\text{ici } \frac{1}{n}\mathbf{W}\right)$ sont les composantes principales du nuage des n points.

Connaissant uniquement les distances d_{ij}, on peut donc calculer les composantes principales, et faire une représentation euclidienne de l'ensemble des points dans un espace de dimension fixée, car les composantes principales ne sont autres que des listes de coordonnées sur une base orthogonale. La dimension de l'espace est alors égale au rang de $\mathbf{W}$: on vérifiera que d est euclidienne si $\mathbf{W}$ a toutes ses valeurs propres positives ou nulles. Remarquons que rang $\mathbf{W} < n - 1$ car n points sont dans un espace de dimension $n - 1$ au plus.

7.5.1.2 Une transformation permettant de passer d'une distance non euclidienne à une distance euclidienne

Si d n'est pas euclidienne, ce qui se produit quand $\mathbf{W}$ a des valeurs propres négatives la méthode de la constante additive permet d'en déduire une distance euclidienne. Il existe en effet une constante c^2, telle que $\delta_{ij}^2 = d_{ij}^2 + c^2$ avec $\delta_{ii} = 0$, soit euclidienne.

La matrice $\mathbf{W}_\delta$ associée à δ est alors telle que :

$$\mathbf{W}_\delta = \mathbf{W}_d + \mathbf{W}_c$$

$$\mathbf{W}_c = -\frac{1}{2}\mathbf{A}\begin{bmatrix} 0 & c^2 & c^2 & c^2 \\ c^2 & 0 & & \\ \cdot & & & \\ c^2 & & & 0 \end{bmatrix}\mathbf{A} = -\frac{1}{2}\mathbf{A}c^2(\mathbf{11}' - \mathbf{I})\mathbf{A}$$

comme $\mathbf{A} = \mathbf{I} - \dfrac{\mathbf{11}'}{n}$, $\mathbf{W}_c = -\dfrac{c^2\mathbf{A}}{2}((n-1)\mathbf{I} - n\mathbf{A})\mathbf{A} = -\dfrac{c^2}{2}((n-1)\mathbf{A} - n\mathbf{A})\mathbf{A} = \dfrac{c^2}{2}\mathbf{A}$
car $\mathbf{A}^2 = \mathbf{A}$.

Les vecteurs propres associés à des valeurs propres non nulles de $\mathbf{W}_d$ sont centrés. Comme A est l'opérateur de centrage, ils sont vecteurs propres de $\mathbf{W}_c$ avec pour valeur propre $c^2/2$.

Aux vecteurs propres de $\mathbf{W}_d$ correspondent les vecteurs propres de $\mathbf{W}_\delta$ avec pour valeurs propres $\lambda + c^2/2$. Il suffit donc de prendre $c^2 = 2|\lambda_n|$ où λ_n est la plus petite valeur propre de $\mathbf{W}_d$ (ici négative) pour que δ soit euclidienne.

Remarquons que cette méthode permet également de transformer une dissimilarité (pour laquelle l'inégalité triangulaire n'est pas vérifiée) directement en une distance euclidienne mais sans doute au prix d'une déformation importante des données.

La méthode précédente (la plus connue) ajoute donc une constante aux carrés des distances. F. Cailliez a résolu en 1983 le problème consistant à ajouter la plus petite constante à la distance d'origine : cette constante est la plus grande valeur propre de la matrice carrée suivante de taille $2n$ $\begin{pmatrix} \mathbf{0} & 2W_d \\ -\mathbf{I} & -4W_{\sqrt{d}} \end{pmatrix}$ où $W_{\sqrt{d}}$ est la matrice de Torgerson où les carrés sont remplacés par les distances.

7.5.2 Le « MDS »

7.5.2.1 Analyse d'un tableau de dissimilarités

Lorsque les d_{ij} ne sont pas des distances mais seulement des mesures de proximité où l'information est de nature ordinale, il est souvent préférable d'utiliser des méthodes semi-métriques de positionnement (***multidimensional scaling***) qui consistent à rechercher une configuration de n points dans un espace euclidien de dimension fixée telle que les distances δ entre ces points respectent au mieux l'ordre défini par d : si $d_{ij} < d_{kl}$, on cherche à avoir $\delta_{ij} < \delta_{kl}$ pour le maximum de points.

Dans l'algorithme MDSCAL de J. B. Kruskal, on cherche à minimiser la quantité suivante appelée *stress* :

$$\min_{\mathbf{e}_i, M} \frac{\sum_{i,j}(\delta_{ij} - M(d_{ij}))^2}{\sum_{i,j}(\delta_{ij})^2}$$

où M est une application monotone croissante.

La méthode est alors la suivante : on part d'une configuration euclidienne obtenue par exemple à l'aide de la formule de Torgerson avec constante additive et on cherche alors

les $M(d_{ij})$ tels que $\sum_{i,j}(\delta_{ij} - M(d_{ij}))^2$ soit minimum. Ce problème admet une solution unique (régression monotone) et on en déduit une valeur du *stress*. On modifie ensuite la configuration au moyen de petits déplacements des points selon une méthode de gradient pour diminuer le *stress*. On repasse ensuite à la phase de régression monotone, etc., jusqu'à convergence.

Une différence fondamentale avec l'analyse d'un tableau de distance euclidienne par ACP est que la dimension de l'espace de représentation doit être fixée à l'avance et que les solutions ne sont pas emboîtées : la meilleure représentation à trois dimensions ne se déduit pas de la meilleure représentation à deux dimensions en rajoutant un troisième axe. Par ailleurs les distances dans l'analyse de Torgerson sont toujours approximées "par en dessous" car la projection raccourcit les distances. La solution du MDS est définie à une transformation orthogonale près (rotation, symétrie, etc.).

7.5.2.2 Analyse de plusieurs tableaux de distances

Pour les mêmes n individus on dispose de q tableaux de distances ou de dissimilarités (par exemple q juges donnent leurs appréciations sur les mêmes objets). Le modèle INDSCAL développé par J.D. Carroll permet de donner une configuration unique des n points et d'étudier les différences entre juges. On se ramène tout d'abord à q matrices de distances euclidiennes par la méthode de la constante additive $\Delta^1, \Delta^2, \ldots, \Delta^q$ on note $d_{ij}^{(k)}$ la distance entre les objets i et j pour le tableau k.

Le modèle INDSCAL postule que :

$$(d_{ij}^{(k)})^2 \# \sum_{l=1}^{r} m_l^{(k)} (x_i^l - x_j^l)^2$$

En d'autres termes il existe une configuration dans un espace à r dimensions pour les objets (coordonnées x_i^l), les juges utilisant des métriques diagonales différentes :

$$\mathbf{M}^{(k)} = \begin{bmatrix} m_1^{(k)} & & 0 \\ & \ddots & \\ 0 & & m_r^{(k)} \end{bmatrix}$$

c'est-à-dire pondérant différemment les dimensions sous-jacentes. Il s'agit donc de trouver une dimension r, les métriques $\mathbf{M}^{(k)}$ et la configuration $\mathbf{X}$ approchant le mieux les données $\Delta^{(k)}$. On convertit tout d'abord les tableaux $\Delta^{(k)}$ en tableaux $\mathbf{W}^{(k)}$ de produits scalaires par la formule de Torgerson et on pose :

$$w_{ij}^k = \sum_{l=1}^{r} m_l^{(k)} a_i^l b_j^l + \varepsilon$$

Si les m et les a sont connus on estime les b par les moindres carrés. Ensuite on estime les m en fixant a et b, puis les a en fixant les m et les b, etc. Les propriétés de symétrie des tableaux $\mathbf{W}^{(k)}$ impliquent que les a et les b sont cohérents ($a_i^l = b_i^l$) et l'algorithme converge. Rien ne prouve cependant que les $m_l^{(k)}$ obtenus soient positifs mais l'expérience montre qu'il en est ainsi dans la plupart des cas avec r faible.

7.6 EXTENSIONS NON LINÉAIRES

L'ACP est une méthode linéaire au sens où les composantes principales sont des combinaisons linéaires des variables initiales et aussi parce qu'elle est basée sur les coefficients de corrélation linéaire. Si les relations entre variables ne sont pas linéaires, l'ACP échoue en général à représenter correctement les données et à extraire de nouvelles variables intéressantes. On sait en effet que le coefficient de corrélation linéaire peut être faible en présence de liaisons fortement non linéaires (*cf.* chapitre 6). La solution est alors de transformer les variables préalablement à l'ACP, afin d'obtenir des corrélations plus élevées et se rapprocher de la linéarité, ce qui revient à se rapprocher de la normalité (*cf.* chapitre 3 page 84).

7.6.1 Recherche de transformations séparées

Il est bien sur possible et souvent recommandé d'utiliser des transformations analytiques classiques (logarithme, puissance, etc.), mais elles ne sont pas forcément optimales. Cette notion d'optimum doit être précisée : on cherchera en général à maximiser le pourcentage d'inertie expliquée par les q premiers axes. La plupart du temps $q = 2$, mais $q = 1$ correspond à des solutions intéressantes.

Pour une variable donnée x_j l'ensemble des transformations $\Phi_j(x_j)$ régulières (au sens de carré intégrable) est bien trop vaste : il est de dimension infinie et conduit à des solutions indéterminées si n est fini, même en imposant que la variable transformée soit centrée-réduite[1]. On se restreindra à des ensembles de transformations correspondant à des espaces vectoriels de dimension finie. Les transformations polynomiales de degré fixé conviendraient mais ont l'inconvénient d'être trop globales et rigides. On leur préfère des transformations polynomiales par morceaux appelées ***fonctions splines***.

Soit x une variable définie sur [a, b] et k points intérieurs régulièrement espacés ou non, on appelle spline de degré d à k nœuds une fonction $S(x)$ qui sur chacun des $k + 1$ intervalles est un polynôme de degré d et est $d - 1$ fois dérivable si $d > 1$, ou seulement continue si $d = 1$ (linéaire par morceaux).

Les splines de degré 2 ou 3 sont souvent utilisées pour leur aspect « lisse ».

Les splines permettent d'approcher toute fonction régulière.

Il est facile de montrer que les splines de degré d à k nœuds forment un espace vectoriel de dimension $d + k + 1$. Tout d'abord les combinaisons linéaires de splines de degré d à k nœuds sont encore des splines de degré d à k nœuds. Sur l'intervalle I_1, le polynôme est libre

1 Dans le cadre de l'ACP entre variables aléatoires (n infini) le problème admet la solution suivante (sans démonstration) liée à l'analyse canonique généralisée de J.D. Carroll. La première composante principale c des variables transformées de façon optimale vérifie $\max_{c,\Phi_j} \sum_{j=1}^{p} \rho^2(c\,;\Phi_j(x^j))$. Pour c fixé $\max_{\Phi_j} \rho^2(c\,;\Phi_j(x^j))$ est atteint pour $\Phi_j(x^j) = E(c/x^j)$. c est donc tel que $\lambda c = \sum_{j=1}^{p} E(c/x^j)$ avec λ maximal.

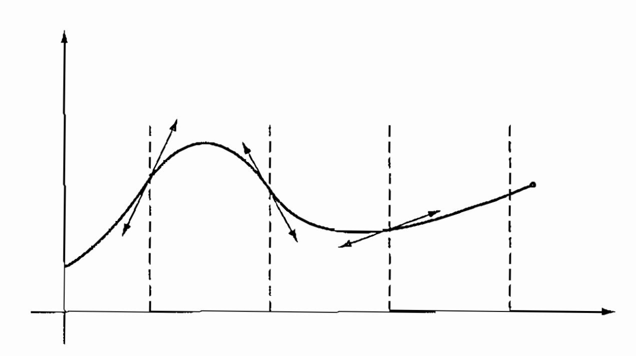


FIGURE 7.8

et dépend de $d + 1$ paramètres, mais sur chacun des k intervalles suivants, les conditions de raccordement (continuité et dérivabilité $d - 1$ fois) ne laissent plus qu'un paramètre libre, d'où le résultat. Puisque l'ensemble des transformations spline est un espace vectoriel, on peut exprimer toute fonction $S(x)$ comme une combinaison linéaire de $d + k + 1$ éléments d'une base, ce qui revient dans un tableau de données X à remplacer chaque colonne-variable par $d + k + 1$ colonnes. On utilisera pour sa simplicité une base permettant des calculs rapides : les B-splines. En voici deux exemples en supposant que a = 0 et b = 1 avec des nœuds régulièrement espacés.

Les splines de degré 0 qui correspondent à des transformations constantes par morceaux (fonctions en escalier) :

$$\begin{cases} B_j(x) = 1 \text{ si } x \in I_j \\ B_j(x) = 0 \text{ sinon} \end{cases}$$

La variable x est alors remplacée par un tableau disjonctif.

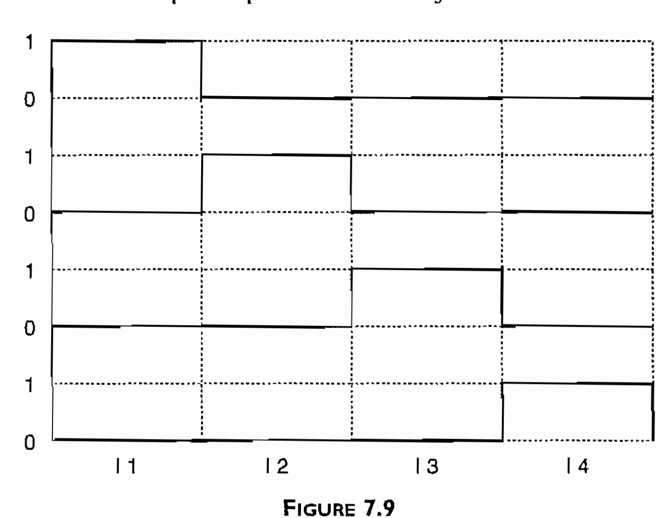


FIGURE 7.9

Les splines de degré 1 correspondent à des transformations continues, linéaires par morceaux. La figure 7.10 donne les cinq fonctions de base associées à trois nœuds.

$$\begin{cases} B_1(x) = 1 - (k+1)x \text{ si } x \in I_1 \\ B_1(x) = 0 \text{ sinon} \end{cases}$$

$$\begin{cases} B_2(x) = (k+1)x \text{ si } x \in I_1 \\ B_2(x) = 2 - (k+1)x \text{ si } x \in I_2 \\ B_2(x) = 0 \text{ sinon} \end{cases}$$

$$\begin{cases} B_{j+1}(x) = (k+1)x - (j-1) \text{ si } x \in I_j \\ B_{j+1}(x) = j + 1 - (k+1)x \text{ si } x \in I_{j+1} \\ B_{j+1}(x) = 0 \text{ sinon} \end{cases}$$

$$\begin{cases} B_{k+2}(x) = (k+1)x - k \text{ si } x \in I_{k+1} \\ B_{k+2}(x) = 0 \text{ sinon} \end{cases}$$

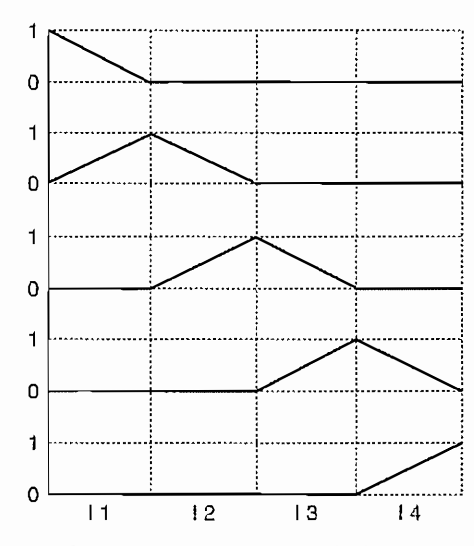


FIGURE 7.10 Les splines de degré 1

Nous ne donnerons pas les formules pour les degrés supérieurs, car de nombreux logiciels les intègrent. Étant donné un point x seules $d + 1$ fonctions de base sont non nulles et de somme égale à 1 (codage « flou »).

La recherche de la transformation de chaque variable ou combinaison linéaire des B-splines, maximisant l'inertie du premier axe de l'ACP s'obtient en effectuant simplement l'ACP du tableau augmenté à n lignes et $p(d + k + 1)$ colonnes. La maximisation de la somme des inerties sur q axes requiert un algorithme plus complexe que nous ne détaillerons pas ici.

Les tranformations splines ne sont pas monotones : on peut aisément y remédier si l'on veut des transformations bijectives. Les B-splines étant positives leurs primitives sont alors des fonctions splines croissantes de degré augmenté d'une unité (I-splines) ; on effectuera alors des combinaisons linéaires à coefficients positifs (*cf.* J.O. Ramsay, 1988).

7.6.2 La « kernel-ACP »

Cette méthode récente (B. Schölkopf *et al.*, 1996) consiste à chercher non plus des transformations séparées de chaque variable mais à transformer tout le vecteur $\mathbf{x} = (x^1, x^2, \ldots, x^p)$. Chaque point de E est alors envoyé dans un espace $\Phi(E)$ muni d'un produit scalaire. La dimension de $\Phi(E)$ peut être très grande et la notion de variable se perd. On effectue alors une analyse factorielle sur tableau de distances entre points transformés selon la méthode de Torgerson qui revient à l'ACP dans $\Phi(E)$. Tout repose sur le choix du produit scalaire dans $\Phi(E)$: si l'on prend un produit scalaire qui s'exprime aisément en fonction du produit scalaire de E, il n'est plus nécessaire de connaître la transformation Φ qui est alors implicite. Tous les calculs s'effectuent en dimension n.

Soit $k(x,y)$ un produit scalaire dans Φ(E) et <x,y> celui de E. Les choix suivants sont couramment utilisés :

$$k(x, y) = (\langle x, y\rangle + c)^d$$

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

$$k(x, y) = \tanh(\langle x, y\rangle + c)$$

Il suffit alors de remplacer la matrice W usuelle par celle où chaque terme est $k(x, y)$, de la centrer en lignes et colonnes et d'en extraire les vecteurs propres pour obtenir les composantes principales dans Φ(E).

Pour que $k(x,y)$ soit bien un produit scalaire, on doit vérifier les conditions de Mercer qui signifient que toute matrice symétrique de terme $k(x,y)$ doit avoir des valeurs propres positives ou nulles.

8 L'analyse canonique et la comparaison de groupes de variables

8.1 ANALYSE CANONIQUE POUR DEUX GROUPES

Lorsque n individus sont décrits par deux ensembles de variables (en nombre p et q respectivement) on cherche à examiner les liens existant entre ces deux ensembles afin de savoir s'ils mesurent ou non les mêmes propriétés.

Exemples : Les deux groupes de notes des disciplines littéraires et scientifiques ; des résultats d'analyses médicales faites par deux laboratoires différents.

Le tableau de données analysé est donc de la forme suivante :

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \left[\begin{array}{c|c} 1 \quad 2 \quad \ldots \quad p & 1 \quad 2 \quad \ldots \quad q \\ & \\ \mathbf{X}_1 & \mathbf{X}_2 \\ & \end{array} \right]$$

On considère alors les deux sous-espaces de $\mathbb{R}^n$ engendrés par les colonnes de $\mathbf{X}_1$ et $\mathbf{X}_2$ respectivement :

$$W_1 = \{\mathbf{x} | \mathbf{x} = \mathbf{X}_1\mathbf{a}\} \qquad \text{et} \qquad W_2 = \{\mathbf{y} | \mathbf{y} = \mathbf{X}_2\mathbf{b}\}$$

W_1 et W_2 sont les deux ensembles de variables que l'on peut construire par combinaisons linéaires des variables de deux groupes. Ces deux espaces peuvent être appelés « potentiels de prévision » (Cailliez, Pagès, 1976).

Si ces deux espaces sont confondus cela prouve que l'on peut se contenter d'un seul des deux ensembles de variables, car ils ont alors même pouvoir de description ; s'ils sont orthogonaux, c'est que les deux ensembles de variables appréhendent des phénomènes totalement différents. Ces deux cas extrêmes étant exceptionnels, on étudiera les positions géométriques de W_1 et W_2 en cherchant les éléments les plus proches, ce qui permettra en particulier de connaître $\dim(W_1 \cap W_2)$.

Si les applications directes de l'analyse canonique sont peu nombreuses, elle n'en constitue pas moins une méthode fondamentale car sa démarche (rechercher des couples de variables en corrélation maximale) se retrouve dans d'autres méthodes comme l'analyse des correspondances, la régression multiple, l'analyse discriminante : si la dimension q de l'un des groupes de variables est égale à 1, l'analyse canonique est équivalente à la régression linéaire multiple étudiée au chapitre 17. Si un des groupes est composé des q variables indicatrices d'une variable qualitative (données réparties en q catégories) et l'autre de p variables numériques, l'analyse canonique conduit à l'analyse factorielle discriminante présentée au chapitre 18. Si les deux groupes des variables sont composés respectivement des indicatrices de deux variables qualitatives à p et q catégories, on obtient l'analyse des correspondances présentée au chapitre 9. ■

8.1.1 Recherche des variables canoniques

On supposera que $\mathbb{R}^n$ est muni de la métrique $\mathbf{D}$. La technique est alors la suivante : chercher le couple $(\boldsymbol{\xi}_1, \boldsymbol{\eta}_1)$ de vecteurs normés où $\boldsymbol{\xi}_1 \in W_1$ et $\boldsymbol{\eta}_1 \in W_2$ forment l'angle le plus faible ; $\boldsymbol{\xi}_1$ et $\boldsymbol{\eta}_1$ sont des combinaisons linéaires respectives des variables du premier et du second groupe appelées variables canoniques.

On recherche ensuite un couple $(\boldsymbol{\xi}_2, \boldsymbol{\eta}_2)$ avec $\boldsymbol{\xi}_2$ $\mathbf{D}$-orthogonal à $\boldsymbol{\xi}_1$ et $\boldsymbol{\eta}_2$ $\mathbf{D}$-orthogonal à $\boldsymbol{\eta}_1$ tels que leur angle soit minimal et ainsi de suite. On obtient ainsi les p couples de variables canoniques (on posera $p = \dim W_1$ et $q = \dim W_2$ avec $p \leq q$).

Notons $\mathbf{A}_1$ et $\mathbf{A}_2$ les opérateurs de projection $\mathbf{D}$-orthogonale sur W_1 et W_2 respectivement.

Il est facile de vérifier que les expressions matricielles explicites de $\mathbf{A}_1$ et $\mathbf{A}_2$ sont (si $\dim W_1 = p$ et $\dim W_2 = q$) :

$$\boxed{\begin{aligned} \mathbf{A}_1 &= \mathbf{X}_1(\mathbf{X}_1'\mathbf{D}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{D} \\ \mathbf{A}_2 &= \mathbf{X}_2(\mathbf{X}_2'\mathbf{D}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{D} \end{aligned}}$$

8.1.1.1 *Étude de la solution dans* $\mathbb{R}^n$

Il s'agit de rechercher deux vecteurs $\boldsymbol{\xi}_1$ et $\boldsymbol{\eta}_1$ de W_1, tels que $\cos(\boldsymbol{\eta}_1, \boldsymbol{\xi}_1)$ soit maximal. En supposant pour l'instant que $\boldsymbol{\eta}_1$ et $\boldsymbol{\xi}_1$ ne sont pas confondus, on voit géométriquement

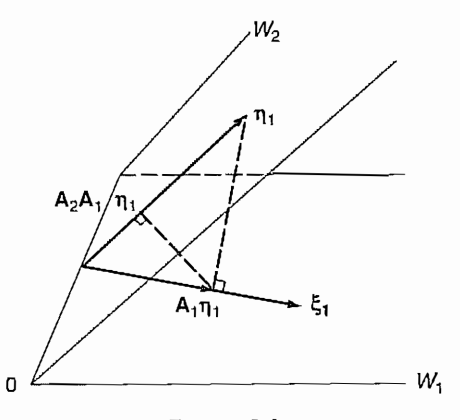


FIGURE 8.1

que η_1 doit être tel que $A_1\eta_1$ sa projection sur W_1 soit colinéaire à ξ_1. En effet, l'élément le plus proche de η_1 est la projection D-orthogonale de η_1 sur W_1. Réciproquement, η_1 doit être l'élément de W_2 le plus proche de ξ_1 (ou de $A_1\eta_1$), donc η_1 doit être colinéaire à $A_2A_1\eta_1$.

Notre problème revient donc à trouver les valeurs propres et les vecteurs propres de A_2A_1 puisque $A_2A_1\eta_1 = \lambda_1\eta_1$.

Inversement, il est immédiat que ξ_1 est vecteur propre de A_1A_2 associé à la même valeur propre.

λ_1 représente le carré du cosinus de l'angle formé par η_1 et ξ_1, ce qui entraîne $\lambda_1 \leq 1$.

Le cas $\lambda_1 = 1$ nous donne $\xi_1 = \eta_1$, donc $\eta_1 \in W_1 \cap W_2$.

Les vecteurs propres de $A_2 A_1$ appartiennent à W_2 :

- en effet, en prémultipliant $A_2A_1\eta_1 = \lambda_1\eta$ par A_2 on trouve puisque $A_2^2 = A_2$, $A_2A_1\eta = \lambda_1A_2\eta$, donc $A_2\eta_1 = \eta_1$;
- on trouve de même que les vecteurs propres de A_1A_2 appartiennent à W_1.

Montrons que A_2A_1 est diagonalisable : puisque les vecteurs propres de A_2A_1 appartiennent nécessairement à W_2 il suffit d'étudier la restriction de A_2A_1 à W_2.

THÉORÈME

*La restriction de A_2A_1 à W_2 est **D**-symétrique.*

Si nous notons $\langle \mathbf{x} ; \mathbf{y} \rangle$ le produit scalaire associé à la métrique **D** :

$$\langle \mathbf{x} ; \mathbf{y} \rangle = \mathbf{x}'\mathbf{D}\mathbf{y}$$

il faut montrer que quel que soit $\mathbf{x}, \mathbf{y} \in W_2$:

$$\langle \mathbf{x} ; A_2A_1\mathbf{y} \rangle = \langle A_2A_1\mathbf{x} ; \mathbf{y} \rangle$$

on a :

$$\begin{aligned}
\langle \mathbf{x} ; A_2A_1\mathbf{y} \rangle &= \langle A_2\mathbf{x} ; A_1\mathbf{y} \rangle && \text{car } A_2 \text{ est } \mathbf{D}\text{-symétrique} \\
&= \langle \mathbf{x} ; A_1\mathbf{y} \rangle && \text{car } \mathbf{x} \in W_2 \\
&= \langle A_1\mathbf{x} ; \mathbf{y} \rangle && \text{car } A_1 \text{ est } \mathbf{D}\text{-symétrique} \\
&= \langle A_1\mathbf{x} ; A_2\mathbf{y} \rangle && \text{car } \mathbf{y} \in W_2 \\
&= \langle A_2A_1\mathbf{x} ; \mathbf{y} \rangle && \text{car } A_2 \text{ est } \mathbf{D}\text{-symétrique}
\end{aligned}$$

c.q.f.d.

Ceci entraîne que la restriction de A_2A_1 à W_2, et par suite A_2A_1, est diagonalisable, ses vecteurs propres sont D-orthogonaux et ses valeurs propres λ_i sont réelles. De plus, les λ_i sont ≥ 0 car A_2 et A_1 sont des matrices positives.

A_2A_1 possède au plus min (p, q) valeurs propres non identiquement nulles. L'ordre de multiplicité de $\lambda_1 = 1$ est alors la dimension de $W_1 \cap W_2$; les vecteurs propres associés à des valeurs propres nulles de rang inférieur à q engendrent la partie de W_2 D-orthogonale à W_1.

Les vecteurs propres $\boldsymbol{\xi}_i$ et $\boldsymbol{\eta}_i$ **D**-normés de $\mathbf{A}_1\mathbf{A}_2$ et de $\mathbf{A}_2\mathbf{A}_1$ sont associés aux mêmes valeurs propres et vérifient les relations suivantes :

$$\mathbf{A}_2\mathbf{A}_1\boldsymbol{\eta}_i = \lambda_i\boldsymbol{\eta}_i \qquad \sqrt{\lambda_i}\boldsymbol{\eta}_i = \mathbf{A}_2\boldsymbol{\xi}_i$$

$$\mathbf{A}_1\mathbf{A}_2\boldsymbol{\xi}_i = \lambda_i\boldsymbol{\xi}_i \qquad \sqrt{\lambda_i}\boldsymbol{\xi}_i = \mathbf{A}_1\boldsymbol{\eta}_i$$

$$\boldsymbol{\eta}'_i\mathbf{D}\boldsymbol{\eta}_j = 0 \quad \text{et} \quad \boldsymbol{\xi}'_i\mathbf{D}\boldsymbol{\xi}_j = 0 \quad \text{pour } i \neq j$$

qui entraînent de plus :

$$\boldsymbol{\eta}'_i\mathbf{D}\boldsymbol{\xi}_j = 0 \quad \text{pour } i \neq j$$

8.1.1.2 Solutions dans $\mathbb{R}^p$ et $\mathbb{R}^q$

Les variables canoniques $\boldsymbol{\xi}_i$ et $\boldsymbol{\eta}_i$ s'expriment comme combinaisons linéaires des colonnes de $\mathbf{X}_1$ et $\mathbf{X}_2$ respectivement :

$$\boldsymbol{\xi}_i = \mathbf{X}_1\mathbf{a}_i \quad \text{et} \quad \boldsymbol{\eta}_i = \mathbf{X}_2\mathbf{b}_i$$

Les $\mathbf{a}_i$ et $\mathbf{b}_i$ sont les facteurs canoniques qui s'obtiennent directement de la manière suivante :

$$\mathbf{A}_1\mathbf{A}_2\boldsymbol{\xi}_i = \lambda_i\boldsymbol{\xi}_i \Leftrightarrow \mathbf{A}_1\mathbf{A}_2\mathbf{X}_1\mathbf{a}_1 = \lambda_i\mathbf{X}_1\mathbf{a}_i$$

en remplaçant les projecteurs par leur expression on a :

$$\mathbf{X}_1(\mathbf{X}'_1\mathbf{D}\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{D}\mathbf{X}_2(\mathbf{X}'_2\mathbf{D}\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{D}\mathbf{X}_1\mathbf{a}_i = \lambda_i\mathbf{X}_1\mathbf{a}_i$$

Si le rang de $\mathbf{X}_1$ est égal au nombre de ses colonnes, on peut simplifier de part et d'autre par $\mathbf{X}_1$ (multiplication par $(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$) et on trouve :

$$(\mathbf{X}'_1\mathbf{D}\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{D}\mathbf{X}_2(\mathbf{X}'_2\mathbf{D}\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{D}\mathbf{X}_1\mathbf{a}_i = \lambda_i\mathbf{a}_i$$

et de même : $$(\mathbf{X}'_2\mathbf{D}\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{D}\mathbf{X}_1(\mathbf{X}'_1\mathbf{D}\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{D}\mathbf{X}_2\mathbf{b}_i = \lambda_i\mathbf{b}_i$$

Dans le cas où toutes les variables sont centrées :

$$\mathbf{X}'_1\mathbf{D}\mathbf{1} = \mathbf{X}'_2\mathbf{D}\mathbf{1} = \mathbf{0}$$

les matrices $\mathbf{X}'_i\mathbf{D}\mathbf{X}_j$ s'interprètent comme des matrices de covariance. On note usuellement :

$$\mathbf{V}_{11} = \mathbf{X}'_1\mathbf{D}\mathbf{X}_1 \qquad \mathbf{V}_{12} = \mathbf{X}'_1\mathbf{D}\mathbf{X}_2$$

$$\mathbf{V}_{22} = \mathbf{X}'_2\mathbf{D}\mathbf{X}_2 \qquad \mathbf{V}_{21} = \mathbf{X}'_2\mathbf{D}\mathbf{X}_1 = (\mathbf{V}_{12})'$$

Les équations des facteurs canoniques s'écrivent alors :

$$\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{a}_i = \lambda_i\mathbf{a}_i$$

$$\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{b}_i = \lambda_i\mathbf{b}_i$$

et les valeurs propres λ_i sont les carrés des coefficients de corrélation canonique entre les variables canoniques. Sur le plan pratique, on diagonalisera évidemment la matrice de taille la plus faible.

Comme on a : $\xi_i = \mathbf{X}_1\mathbf{a}_i$ et $\eta_i = \mathbf{X}_2\mathbf{b}_i$ si l'on désire que les variables canoniques soient de variance unité, on normera les facteurs principaux de la manière suivante :

$$\boxed{\mathbf{a}_i'\mathbf{V}_{11}\mathbf{a}_i = 1 \qquad \text{et} \qquad \mathbf{b}_i'\mathbf{V}_{22}\mathbf{b}_i = 1}$$

On en déduit :

$$\boxed{\mathbf{b}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{a}_i \qquad \text{et} \qquad \mathbf{a}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{b}_i}$$

Comme seuls comptent les sous-espaces W_1 et W_2, il est équivalent de travailler avec des variables de variance 1. On utilisera donc en général les matrices de corrélation $\mathbf{R}_{11}$, $\mathbf{R}_{12}$, etc à la place des matrices de variance.

8.1.2 Représentation des variables et des individus

Deux sortes de représentations sont possibles selon qu'on choisit les variables canoniques de W_1 ou de W_2. Si l'on fait choix de W_1 on représentera l'ensemble des variables de départ **D**-normées (colonnes de $\mathbf{X}_1$ et de $\mathbf{X}_2$) en projection sur la base **D**-orthonormée formée par les ξ_i.

En particulier, la projection sur le plan engendré par ξ_1 et ξ_2 donne un cercle des corrélations (fig. 8.2) car, si les colonnes de $\mathbf{X}_1$ sont **D**-normées ainsi que celles de $\mathbf{X}_2$, les composantes sur la base des ξ_i sont les coefficients de corrélation entre les variables initiales et les variables canoniques.

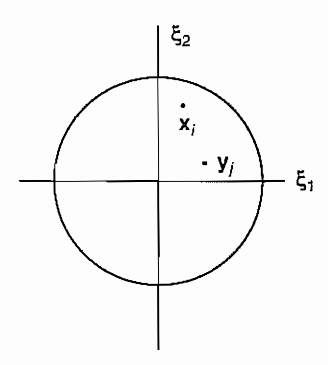


FIGURE 8.2

Si $\mathbf{x}_k$ est la $k^{\text{ième}}$ colonne de $\mathbf{X}_1$ on a $\mathbf{x}_k'\mathbf{D}\xi_1 = \mathbf{x}_k'\mathbf{D}\mathbf{X}_1\mathbf{a}_1$; le coefficient de corrélation entre $\mathbf{x}_k$ et ξ_1 est la $k^{\text{ième}}$ composante de $\mathbf{V}_{11}\,\mathbf{a}_1$ car $\mathbf{x}_k$ est égal à $\mathbf{X}_1\boldsymbol{\delta}_k$ où $\boldsymbol{\delta}_k$ est le vecteur de $\mathbb{R}^p$ dont toutes les composantes sont nulles sauf la $k^{\text{ième}}$ qui vaut 1.

Si $\mathbf{y}_l$ est la $l^{\text{ième}}$ colonne de $\mathbf{X}_2$:

$$\mathbf{y}_l'\,\mathbf{D}\xi_1 = \boldsymbol{\delta}_l'\,\mathbf{X}_2'\,\mathbf{D}\mathbf{X}_1\mathbf{a}_l$$

la corrélation entre $\mathbf{y}_l$ et ξ_i est alors la $l^{\text{ième}}$ composante de $\mathbf{V}_{21}\mathbf{a}_i$ ou encore la $l^{\text{ième}}$ composante de $\sqrt{\lambda_1}\mathbf{V}_{22}\mathbf{b}_i$.

Si les colonnes de $\mathbf{X}_1$ et $\mathbf{X}_2$ ne sont pas **D**-normées il faut diviser les expressions précédentes par les normes de $\mathbf{x}_k$ ou $\mathbf{y}_l$.

Les représentations sur (ξ_1, ξ_2) et (η_1, η_2) sont d'autant plus voisines que λ_1 et λ_2 sont proches de 1.

Pour les individus deux représentations des individus sont possibles selon les variables canoniques choisies.

Si l'on choisit le plan défini par (ξ_1, ξ_2) les coordonnées du $j^{\text{ième}}$ point sont les $j^{\text{ième}}$ composantes des variables canoniques ξ_1 et ξ_2.

8.1.3 Test du nombre de variables canoniques significatives

On peut arrêter l'extraction des valeurs propres et des vecteurs propres au moyen du test de Bartlett.

L'hypothèse que les deux ensembles de variables sont indépendants revient à tester $\lambda_1 = 0$. Si cette hypothèse est rejetée, on teste la nullité de λ_2, etc.

D'une façon générale, si $\lambda_1, \lambda_2, \ldots, \lambda_k$ sont jugés significativement différents de zéro, on teste la nullité des valeurs propres suivantes en utilisant la quantité :

$$-\left[n - 1 - k - \frac{1}{2}(p + q + 1) + \sum_{i=1}^{k}\frac{1}{\lambda_i}\right] \ln\left(\prod_{k+1}^{\min(p,q)}(1 - \lambda_i)\right)$$

qui suit approximativement un $\chi^2_{(p-k)(q-k)}$, si la valeur théorique de λ_{k+1} (donc de $\lambda_{k+2}\ldots$) est nulle.

Le test précédent n'est valide que dans le cas de variables normales et ne s'applique en aucune façon aux cas particuliers que sont l'analyse des correspondances et l'analyse discriminante.

8.2 MÉTHODES NON SYMÉTRIQUES POUR DEUX GROUPES DE VARIABLES

L'analyse canonique est une méthode faisant jouer des rôles symétriques aux deux groupes de variables. Si l'un d'entre eux est privilégié diverses approches sont possibles.

8.2.1 Méthodes procustéennes de comparaison de deux configurations d'individus

On suppose ici que les deux groupes de variables ont même dimension (cas auquel on peut toujours se ramener en rajoutant des coordonnées nulles) afin de confondre les espaces W_1 et W_2.

On dispose donc de deux cartes p-dimensionnelles des mêmes n individus obtenues par des procédés différents et on cherche à les comparer.

Le principe consiste alors à fixer l'une des deux configurations (le tableau $\mathbf{X}_1$) et à faire subir à l'autre une transformation géométrique simple ($\tilde{\mathbf{X}}_2 = \mathbf{X}_2\mathbf{T}$) telle que les deux configurations deviennent les plus proches possibles, un critère naturel étant :

$$\min_{\mathbf{T}} \sum_{i=1}^{n} \|\mathbf{e}_i - \tilde{\mathbf{e}}_i\|^2 = \min_{\mathbf{T}} \text{ Trace }[(\mathbf{X}_1 - \mathbf{X}_2\mathbf{T})(\mathbf{X}_1 - \mathbf{X}_2\mathbf{T})']$$

Ce type de problème se rencontre en particulier pour comparer des solutions différentes de *multidimensional scaling* où les axes ne sont pas les mêmes.

On suppose dans la suite que les poids des individus sont égaux à $1/n$, mais il est facile d'étendre les résultats au cas général.

Si $\mathbf{T}$ est une transformation quelconque la solution des moindres carrés est donnée par :

$$\mathbf{T} = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1 = \mathbf{V}_{22}^{-1}\mathbf{V}_{21}$$

En général, on impose à $\mathbf{T}$ d'être une transformation orthogonale :

l'idée est de faire subir à l'une des deux configurations une série de rotations, symétries, retournements de façon à l'amener le plus possible sur l'autre.

Le problème devient :

$$\min_{\mathbf{T}} \text{ Trace } [(\mathbf{X}_1 - \mathbf{X}_2\mathbf{T})(\mathbf{X}_1 - \mathbf{X}_2\mathbf{T})'] \qquad \text{avec} \qquad \mathbf{TT}' = \mathbf{I}$$

soit $\dfrac{p(p+1)}{2}$ contraintes.

Réécrivons tout d'abord le critère :

Trace $(\mathbf{X}_1 - \mathbf{X}_2\mathbf{T})(\mathbf{X}_1 - \mathbf{X}_2\mathbf{T})' =$ Trace $\mathbf{X}_1\mathbf{X}_1'$ + Trace $\mathbf{X}_2\mathbf{TT}'\mathbf{X}_2'$ − 2 Trace $\mathbf{X}_1'\mathbf{X}_2\mathbf{T}$. Si $\mathbf{TT}' = \mathbf{I}$ on voit que $\mathbf{T}$ doit maximiser Trace $\mathbf{V}_{12}\mathbf{T}$ sous la contrainte $\mathbf{TT}' = \mathbf{I}$.

Introduisons alors la matrice $\Lambda/2$ symétrique de taille p des multiplicateurs de Lagrange associés aux $\dfrac{p(p+1)}{2}$ contraintes. On doit alors rendre maximum :

$$\text{Trace}\left[\mathbf{V}_{12}\mathbf{T} - \frac{1}{2}\Lambda(\mathbf{TT}' - \mathbf{I})\right]$$

en dérivant cette expression par rapport à la matrice $\mathbf{T}$ on obtient le système d'équations :

$$\mathbf{V}_{21} = \Lambda\mathbf{T} \qquad \text{soit} \qquad \Lambda = \mathbf{V}_{21}\mathbf{T}' \qquad \text{en multipliant par } \mathbf{T}'$$

car $\dfrac{d}{d\mathbf{T}}$ Trace $\mathbf{V}_{12}\mathbf{T} = \mathbf{V}_{21}$ et $\dfrac{d}{d\mathbf{T}}$ Trace $\Lambda\mathbf{TT}' = 2\Lambda\mathbf{T}$.

Pour trouver $\mathbf{T}$ nous écrivons $\mathbf{V}_{21}$ sous forme de décomposition en valeurs singuliéres.

$\mathbf{V}_{21} = \mathbf{VSU}'$ où $\mathbf{S}$ est la matrice diagonale des valeurs propres de $\mathbf{V}_{21}\mathbf{V}_{12}$, $\mathbf{V}$ la matrice orthogonale des vecteurs propres normés de $\mathbf{V}_{21}\mathbf{V}_{12}$, $\mathbf{U}$ la matrice orthogonale des vecteurs propres normés de $\mathbf{V}_{12}\mathbf{V}_{21}$.

On en déduit :

$$\Lambda = \mathbf{VSU}'\mathbf{T}' = \mathbf{TUSV}' \qquad \text{car} \qquad \Lambda \text{ est symétrique}$$

d'où $\Lambda^2 = \mathbf{VSU}'\mathbf{T}'\mathbf{TUSV}' = \mathbf{VS}^2\mathbf{V}'$ donc $\Lambda = \mathbf{VSV}'$ et $\mathbf{V}_{21} = \Lambda\mathbf{T}$ donne $\mathbf{VSU}' = \mathbf{VSV}'\mathbf{T}$

La meilleure transformation orthogonale $\mathbf{T}$ est donc telle que :

$$\boxed{\mathbf{T} = \mathbf{VU}'}$$

8.2.2 Méthodes factorielles

Leur principe consiste à chercher des combinaisons linéaires de variables d'un des deux groupes vérifiant certaines conditions ou contraintes liées à l'existence du deuxième groupe de variables. Selon que l'on cherche à se rapprocher du deuxième groupe ou au contraire à s'affranchir de son influence on pourra utiliser :

8.2.2.1 L'analyse en composantes principales de variables instrumentales (ACPVI)

On recherche des combinaisons linéaires ξ des variables du premier groupe « expliquant » le mieux les variables du deuxième groupe. C. R. Rao a introduit le critère suivant :

" Si l'on régresse les m_2 variables du tableau $\mathbf{X}_2$ sur ξ, la somme des variances résiduelles doit être minimale. "

En posant $\xi = \mathbf{X}_1\mathbf{a}$, ce critère revient à rendre maximale la somme des variances expliquées soit à un coefficient près :

$$\sum_{j=1}^{m_2}(\mathbf{x}_2^j)'\,\xi(\xi'\xi)^{-1}\,\xi'(\mathbf{x}_2^j)$$

$$= \sum_{j=1}^{m_2}\frac{(\mathbf{x}_2^j)'\mathbf{X}_1\mathbf{a}\mathbf{a}'\mathbf{X}_1'(\mathbf{x}_2^j)}{\mathbf{a}'\mathbf{X}_1'\mathbf{X}_1\mathbf{a}} = \text{Trace}\,\frac{\mathbf{X}_2'\mathbf{X}_1\mathbf{a}\mathbf{a}'\mathbf{X}_1'\mathbf{X}_2}{\mathbf{a}'\mathbf{X}_1'\mathbf{X}_1\mathbf{a}}$$

$$= \text{Trace}\,\frac{\mathbf{V}_{21}\mathbf{a}\mathbf{a}'\mathbf{V}_{12}}{\mathbf{a}'\mathbf{V}_{11}\mathbf{a}} = \frac{\mathbf{a}'\mathbf{V}_{12}\mathbf{V}_{21}\mathbf{a}}{\mathbf{a}'\mathbf{V}_{11}\mathbf{a}}$$

$\mathbf{a}$ est donc vecteur propre associé à sa plus grande valeur propre, de la matrice :

$$\boxed{\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{21}}$$

Les valeurs propres suivantes conduisent à d'autres solutions non corrélées entre elles.

Lorsque $\mathbf{X}_2$ est un ensemble de variables de variance unité, ξ est la combinaison linéaire des variables de $\mathbf{X}_1$ la plus corrélée avec les variables de $\mathbf{X}_2$ au sens où :

$$\boxed{\sum_{j=1}^{m_2} r^2(\xi\,;\,\mathbf{x}_2^j)\ \text{est maximal}}$$

On reconnaît ici une expression voisine du critère usuel de l'ACP réduite : ici on calcule les corrélations avec des variables externes.

Les variables ξ sont les composantes principales de l'ACP de $\mathbf{X}_1$ avec pour métrique $\mathbf{M} = \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{21}\mathbf{V}_{11}^{-1} = (\mathbf{V}_{11}^{-1}\mathbf{V}_{12})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12})'$ ou, ce qui est équivalent, les composantes principales du tableau $\mathbf{X}_1\mathbf{V}_{11}^{-1}\mathbf{V}_{12}$ avec la métrique identité : en d'autres termes on effectue l'ACP des projections des variables de $\mathbf{X}_2$ sur $\mathbf{X}_1$.

Le coefficient de redondance de Stewart et Love entre deux groupes de variables : $R^2(\mathbf{X}_2 : \mathbf{X}_i)$ (notons que $R^2(\mathbf{X}_1 : \mathbf{X}_2) \neq R^2(\mathbf{X}_2 : \mathbf{X}_1)$) :

$$\text{est :}\ R^2(\mathbf{X}_2 : \mathbf{X}_1) = \frac{\text{Trace}\,(\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})}{\text{Trace}\,\mathbf{V}_{22}}$$

Lorsque $V_{22} = R_{22}$ (variables de X_2 standardisées) $R_2(X_2 : X_1) = \frac{1}{m_2}\sum_{j=1}^{m_2} R^2(x_2^j ; X_1)$ moyenne des carrés des coefficients de corrélation multiple des régressions des x_2^j sur X_1.

On voit alors que les composantes principales des variables instrumentales ξ sont les combinaisons linéaires des colonnes de X_1 ayant une redondance maximale avec X_2.

On vérifie aisément que ξ est vecteur propre de A_1W_2 où :

$$A_1 = X_1(X_1'X_1)^{-1}X_1' \qquad \text{et} \qquad W_2 = X_2X_2'$$

8.2.2.2 *ACP sous contrainte d'orthogonalité*

On peut inversement rechercher des combinaisons linéaires des variables de X_1 de variance maximale sous la contrainte d'être non corrélées aux variables de X_2 afin d'éliminer leur effet. Pour que le problème ait une solution il faut que $m_2 < m_1$. On montre alors que les facteurs **a** tels que $\xi = X_1\mathbf{a}$ sont vecteurs propres de :

$$(\mathbf{I} - V_{12}(V_{21}V_{12})^{-1}V_{21})V_{11}$$

8.2.2.3 *ACP des covariances partielles*

Une autre manière d'éliminer l'influence des variables extérieures X_2 consiste à utiliser la matrice des covariances (ou des corrélations) partielles de X_1 à X_2 fixé :

$$V_{11/2} = V_{11} - V_{12}V_{22}^{-1}V_{21}$$

On cherche alors les vecteurs propres de cette matrice. Il s'agit ici d'une ACP du nuage des résidus des régressions des variables de X_1 sur X_2 : les « composantes principales » ne sont pas ici des combinaisons linéaires des variables de X_1.

8.3 L'ANALYSE CANONIQUE GÉNÉRALISÉE

Étendre l'analyse canonique à plus de deux groupes de variables se heurte d'emblée à la difficulté suivante : il n'existe pas de mesure simple de la liaison entre plus de deux variables. Il y aura donc autant de façons d'obtenir des variables canoniques que de manières de définir une « corrélation » entre p variables : on peut prendre par exemple comme mesure la somme des corrélations deux à deux, la somme des carrés des corrélations, le déterminant de la matrice des corrélations, etc. Toute généralisation est donc plus ou moins arbitraire. Celle que nous présentons ici a l'avantage d'être sans doute la plus simple et la plus riche d'interprétations, car elle se relie aisément à toutes les autres méthodes d'analyse des données.

8.3.1 Une propriété de l'analyse canonique ordinaire

Étant donné deux ensembles de variables centrées X_1 et X_2, les variables canoniques ξ et η, vecteurs propres de A_1A_2 et A_2A_1 respectivement, possèdent la propriété suivante :

$$\xi + \eta \text{ est vecteur propre de } A_1 + A_2$$

En effet, posons **z** tel que $(A_1 + A_2)\mathbf{z} = \mu\mathbf{z}$; en prémultipliant par A_1 ou A_2 cette équation, on trouve en utilisant l' idempotence de A_1 et A_2 :

$$A_1(A_1 + A_2)\mathbf{z} = \mu A_1\mathbf{z}$$

soit : $$A_1A_2\mathbf{z} = (\mu - 1)A_1\mathbf{z} \qquad \text{et} \qquad A_2A_1\mathbf{z} = (\mu - 1)A_2\mathbf{z}$$

ce qui donne :

$$A_1A_2A_1\mathbf{z} = (\mu - 1)^2A_1\mathbf{z}$$
$$A_2A_1A_2\mathbf{z} = (\mu - 1)^2A_2\mathbf{z}$$

donc au même coefficient multiplicateur près, $A_1\mathbf{z}$ et $A_2\mathbf{z}$ ne sont autres que les variables canoniques $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$; comme $A_1\mathbf{z} + A_2\mathbf{z} = \mu\mathbf{z}$ on trouve $\mu\mathbf{z} = \boldsymbol{\xi} + \boldsymbol{\eta}$, ce qui démontre la propriété annoncée (fig. 8.3).

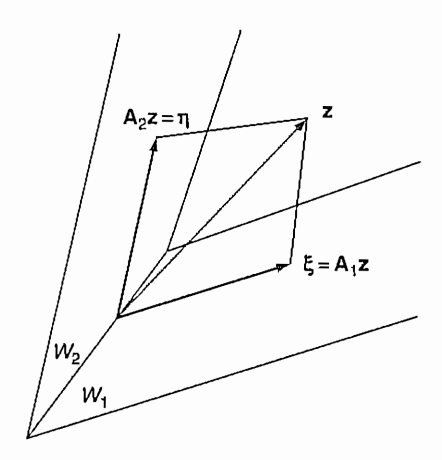


FIGURE 8.3

La variable **z** possède la propriété d'être la plus liée aux deux ensembles $\mathbf{X}_1$ et $\mathbf{X}_2$, en ce sens qu'elle a une somme des carrés des coefficients de corrélation multiple maximale avec $\mathbf{X}_1$ et $\mathbf{X}_2$.

En effet, le coefficient de corrélation multiple de **z** avec $\mathbf{X}_i$ vaut :

$$R_i^2 = \frac{\mathbf{z}'\mathbf{D}A_i\mathbf{z}}{\mathbf{z}'\mathbf{D}\mathbf{z}} = \frac{\|A_i\mathbf{z}\|^2}{\|\mathbf{z}\|^2}$$

car les variables étant centrées, R_i est le cosinus de l'angle formé par **z** et W_i.

8.3.2 La généralisation de J. D. Carroll (1968)

De la propriété précédente découle la généralisation suivante due à J. D. Carroll : plutôt que de rechercher directement des variables canoniques dans chacun des sous-espaces W_i associés à des tableaux de données $\mathbf{X}_i$, on cherche une variable auxiliaire **z** appartenant à la somme des W_i telle que $\sum_{i=1}^{p} R^2(\mathbf{z} ; \mathbf{X}_i)$ soit maximal.

$\mathbf{z}$ est alors vecteur propre de $\mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_p$:

$$\boxed{(\mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_p)\mathbf{z} = \mu\mathbf{z}}$$

On obtient ensuite, si nécessaire, des variables canoniques ξ_i en projetant $\mathbf{z}$ sur les W_i. $\xi_i = \mathbf{A}_i\mathbf{z}$.

Si l'on pose $\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2| \ldots |\mathbf{X}_p)$, matrice à n lignes et $\sum_{i=1}^{p} m_i$ colonnes, la variable $\mathbf{z}$ se met sous la forme $\mathbf{Xb}$ et plutôt que de rechercher $\mathbf{z}$ comme vecteur propre d'une matrice n, n il vaut mieux chercher $\mathbf{b}$ qui possède Σm_i composantes.

Comme $\mathbf{A}_i = \mathbf{X}_i(\mathbf{X}_i'\mathbf{D}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{D}$, en posant $\mathbf{V}_{ii} = \mathbf{X}_i'\mathbf{D}\mathbf{X}_i$ matrice de variance-covariance du $i^{\text{ème}}$ groupe et $\mathbf{M} = \begin{bmatrix} \mathbf{V}_{11}^{-1} & & & \\ & \mathbf{V}_{22}^{-1} & & \\ & & \ddots & \\ & & & \mathbf{V}_{pp}^{-1} \end{bmatrix}$ matrice bloc-diagonale des $\mathbf{V}_{ii}^{-1}$, on trouve aisément que $\sum_{i=1}^{p}\mathbf{A}_i = \sum_{i=1}^{p}\mathbf{X}_i\mathbf{V}_{ii}^{-1}\mathbf{X}_i'\mathbf{D}$ s'écrit en fait $\sum_{i=1}^{p}\mathbf{A}_i = \mathbf{XMX'D}$.

Donc $\mathbf{z}$ est vecteur propre de $\mathbf{XMX'D}$, et puisque $\mathbf{z} = \mathbf{Xb}$, si $\mathbf{X}$ est de plein rang, $\mathbf{b}$ est vecteur propre de $\mathbf{MX'DX}$:

$$\boxed{\begin{aligned} \mathbf{XMX'Dz} &= \mu\mathbf{z} \\ \mathbf{MX'DXb} &= \mu\mathbf{b} \end{aligned}}$$

On reconnaît alors les équations donnant les composantes principales et les facteurs principaux, dans l'ACP du tableau total $\mathbf{X}$ avec la métrique $\mathbf{M}$.

En particulier si chaque groupe est réduit à une seule variable ($m_i = 1, i = 1, 2, \ldots, p$) on retrouve l'ACP avec la métrique $\mathbf{D}_{1/s^2}$ puisque $\mathbf{z}$ rend alors maximal $\sum_{i=1}^{p} r^2(\mathbf{z}\,;\mathbf{x}^i)$.

L'analyse canonique généralisée est donc une ACP sur des groupes de variables, ce qui nous ramène à une optique de description des individus tenant compte des liaisons par blocs plutôt qu'à une optique de description des relations entre variables.

On a toujours $\Sigma\mu_k = \Sigma m_i$ Si $\mu = p$, il existe une intersection commune à tous les W_i.

Les « variables canoniques » $\xi_i^{(k)}$ que l'on déduit des $\mathbf{z}^{(k)}$ par projection orthogonale sur les W_i ont alors la propriété suivante, du moins pour l'ordre 1 : le p-uple $(\xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_p^{(1)})$ a une matrice de corrélation dont la plus grande valeur propre λ_1 est maximale.

Contrairement à l'analyse canonique usuelle avec $p = 2$, il n'y a pas orthogonalité entre les $\xi_i^{(k)}$ et les $\xi_i^{(l)}$.

On verra au chapitre 10 que si chaque groupe est celui des variables indicatrices de p variables qualitatives, l'analyse canonique généralisée conduit à l'analyse des correspondances multiples. L'analyse canonique généralisée de Carroll n'est pas la seule méthode de traitement simultané de p groupes de variables ; de très nombreuses techniques ont été proposées : la méthode Statis, l'analyse factorielle multiple sont les plus connues. Le lecteur intéressé est invité à consulter les ouvrages de B. Escofier, du Geri, ainsi que l'article de synthèse de P. Cazes (2004) cités en bibliographie.

9 L'analyse des correspondances

Cette méthode a été proposée en France par J.-P. Benzécri dans le but d'étudier la liaison (dite encore correspondance) entre deux variables qualitatives ; un exemple de correspondance nous est fourni, par exemple, par la ventilation des séjours de vacances selon le mode d'hébergement et la catégorie socio-professionnelle (CSP) (voir chapitre 6, § 6.5).

Sur le plan mathématique, on peut considérer l'analyse des correspondances soit comme une analyse en composantes principales avec une métrique spéciale, la métrique du χ^2, soit comme une variante de l'analyse canonique. Nous développerons ces deux aspects en accordant toutefois la préférence à l'aspect analyse canonique qui a entre autres avantages de respecter la symétrie entre les deux variables et de généraliser sans difficulté l'analyse des correspondances à plusieurs variables qualitatives.

9.1 TABLEAU DE CONTINGENCE ET NUAGES ASSOCIÉS

9.1.1 Représentations géométriques des profils associés à un tableau de contingence

Le tableau des données est un tableau de contingence N à m_1 lignes et m_2 colonnes résultant du croisement de deux variables qualitatives à m_1 et m_2 catégories respectivement (voir chapitre 6, paragr. 6.5).

Si l'on note D_1 et D_2 les matrices diagonales des effectifs marginaux des deux variables :

$$D_1 = \begin{bmatrix} n_{1.} & & & & 0 \\ & n_{2.} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & n_{m_1} \end{bmatrix} \qquad D_2 = \begin{bmatrix} n_{.1} & & & & 0 \\ & n_{.2} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & n_{.m_2} \end{bmatrix}$$

Le tableau des profils des lignes d'éléments $\dfrac{n_{ij}}{n_{i.}}$ est alors $\mathbf{D}_1^{-1}\,\mathbf{N}$.

Le tableau des profils des colonnes d'éléments $\dfrac{n_{ij}}{n_{.j}}$ est alors $\mathbf{N}\mathbf{D}_2^{-1}$.

Les profils de lignes forment un nuage de m_1 points dans $\mathbb{R}^{m_2}$; chacun de ces points étant affecté d'un poids proportionnel à sa fréquence marginale $\left(\text{matrice de poids} : \dfrac{\mathbf{D}_1}{n}\right)$.

Le centre de gravité de ce nuage de points est :

$$\mathbf{g}_l = \frac{1}{n}(\mathbf{D}_1^{-1}\mathbf{N})'\mathbf{D}_1\mathbf{1} = \begin{bmatrix} \frac{n_{.1}}{n} \\ \frac{n_{.2}}{n} \\ . \\ . \end{bmatrix} = \begin{bmatrix} p_{.1} \\ p_{.2} \\ . \\ p_{.m_2} \end{bmatrix}$$

c'est-à-dire le profil marginal.

Réciproquement, les profils-colonnes forment un nuage de m_2 points dans $\mathbb{R}^{m_1}$ avec des poids donnés par la matrice $\dfrac{\mathbf{D}_2}{n}$; leur centre de gravité $\mathbf{g}_c$ est le point de coordonnées :

$$\mathbf{g}_c = \begin{bmatrix} p_{1.} \\ p_{2.} \\ . \\ . \\ p_{m_1.} \end{bmatrix}$$

Pour garder les conventions du chapitre 7, les profils des colonnes de $\mathbf{N}$ sont les lignes du tableau transposé $\mathbf{D}_2^{-1}\mathbf{N}'$ (« individus » en lignes, « variables » en colonnes).

Dans le cas de l'indépendance statistique :

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \qquad \text{et} \qquad \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$$

les deux nuages sont alors réduits chacun à un point, leurs centres de gravité respectifs.

L'étude de la forme de ces nuages au moyen de l'analyse en composantes principales permettra donc de rendre compte de la structure des écarts à l'indépendance mais il faut choisir alors une métrique pour chacun de ces espaces.

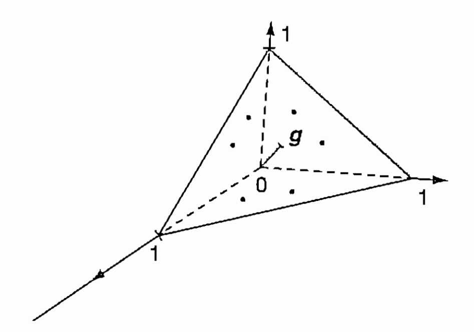


FIGURE 9.1

Remarquons que les profils ayant pour somme 1, les m_1 profils-lignes sont en réalité situés dans le sous-espace W_2 de dimension $m_2 - 1$ défini par $\sum_{j=1}^{m_2} x_j = 1$ (avec en plus $x_j \geq 0$) ainsi que leur centre de gravité (fig. 9.1). De même pour les m_2 profils des colonnes.

9.1.2 La métrique du χ^2

Pour calculer la distance entre deux profils-lignes i et i' on utilise la formule suivante :

$$d^2_{\chi^2}(i, i') = \sum_{j=1}^{m_2} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2$$

Il s'agit donc de la métrique diagonale $n\mathbf{D}_2^{-1}$.

La considération suivante justifie dans une certaine mesure l'emploi de la distance du χ^2 :

La pondération par $\frac{n}{n_{.j}}$ de chaque carré de différence revient à donner des importances comparables aux diverses « variables » : ainsi, dans l'exemple de la correspondance modes d'hébergement × CSP, (voir chapitre 6 et § 9.3) si l'on calculait la distance entre deux modes par la formule usuelle : « somme des carrés des différences des pourcentages des diverses CSP », il est clair que cette distance refléterait surtout la différence entre les CSP les plus importantes en nombre ; pour pallier cet inconvénient la division par $n_{.j}$ est un bon remède (quoiqu'un peu arbitraire).

L'argument le plus fréquemment utilisé en faveur de la métrique du χ^2 est le principe d'équivalence distributionnelle : si deux colonnes de $\mathbf{N}$, j et j', ont même profil il est logique de les regrouper en une seule d'effectifs $(n_{ij} + n_{ij'})$, il faut alors que cette opération ne modifie pas les distances entre profils-lignes.

On vérifie en effet par simple calcul que :

$$\frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 + \frac{n}{n_{.j'}} \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}} \right)^2 = \frac{n}{n_{.j} + n_{.j'}} \left(\frac{n_{ij} + n_{ij'}}{n_{i.}} - \frac{n_{i'j} + n_{i'j'}}{n_{i'.}} \right)^2$$

lorsque $\frac{n_{ij}}{n_{.j}} = \frac{n_{ij'}}{n_{.j'}}$.

Cette propriété n'est pas vérifiée pour la métrique euclidienne usuelle.

La justification la plus profonde, mais la plus difficile, est en fait la suivante : les profils-lignes sont des lois de probabilité sur des ensembles finis de m_2 éléments (les modalités de la deuxième variable). Au moyen de l'espérance mathématique, à ces lois de probabilité sont associées des formes linéaires (qu'on leur identifie) sur les variables quantitatives compatibles avec la deuxième variable qualitative. Ces variables quantitatives (qui réalisent une quantification de la deuxième variable qualitative) formant un espace vectoriel, les « individus » sont donc des éléments du dual de cet espace (pas tout le dual, mais un simplexe de ce dual).

Les modalités de la deuxième variable ayant pour poids $p_{.1}, p_{.2}, \ldots,$ les variables quantitatives associées sont munies de la métrique $\frac{1}{n}\mathbf{D}_2$ qui est la métrique de la covariance,

si l'on se restreint à des codages centrés. Le dual doit donc être muni de la métrique inverse $n\mathbf{D}_2^{-1}$.

On définit de même la métrique du χ^2 entre profils-colonnes (matrice $n\mathbf{D}_1^{-1}$) par la formule :

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{m_1} \frac{n}{n_{i.}}\left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right)^2$$

Le terme de métrique du χ^2 vient de ce que les deux nuages ont alors pour inertie totale la quantité mesurant l'écart à l'indépendance :

$$\varphi^2 = \frac{1}{n}\sum_i \sum_j \frac{\left(n_{ij} - \dfrac{n_{i.}\,n_{.j}}{n}\right)^2}{\dfrac{n_{i.}\,n_{.j}}{n}} \qquad \text{(voir chapitre 6)}$$

En effet, l'inertie du nuage des profils-lignes par rapport à $\mathbf{g}_l$ vaut :

$$\sum_{i=1}^{m_1} \frac{n_{i.}}{n} d^2_{\chi^2}(\mathbf{i}, \mathbf{g}_l) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{n_{i.}}{n}\frac{n}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}\right)^2$$

ce qui donne φ^2 après un calcul élémentaire. Il en est de même pour l'inertie du nuage des profils-colonnes.

Nous avons remarqué que le nuage des points profils-lignes était dans un sous-espace W_1 : le vecteur $\mathbf{Og}_l$ est alors orthogonal au sens de la métrique du χ^2 à ce sous-espace (fig. 9.2) :

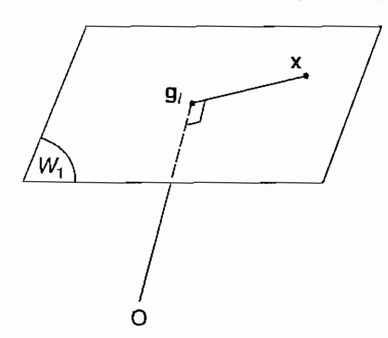


FIGURE 9.2

En effet, soit $\mathbf{x}$ un élément de W_1 :

$$(\mathbf{x} - \mathbf{g}_l)' n\mathbf{D}_2^{-1}\mathbf{g}_l = \langle \mathbf{Og}_l \,;\, \mathbf{g}_l\mathbf{x} \rangle_{\chi^2} = 0$$

car :

$$n\mathbf{D}_2^{-1}\mathbf{g}_l = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

et pour tout élément de W_1 : $\mathbf{x}'\mathbf{1} = 1$ donc $\mathbf{g}_l'\,\mathbf{1} = 1$.

De plus $\|\mathbf{g}_l\|^2_{\chi^2} = \mathbf{g}_l'\mathbf{1} = 1$.

9.2 ANALYSES EN COMPOSANTES PRINCIPALES DES DEUX NUAGES DE PROFILS

Deux ACP sont alors possibles :

1) Celle du nuage des profils-lignes avec :
   - tableau de données $\mathbf{X} = \mathbf{D}_1^{-1}\mathbf{N}$;
   - métrique $\mathbf{M} = n\mathbf{D}_2^{-1}$;
   - poids $\mathbf{D} = \dfrac{\mathbf{D}_1}{n}$.

2) Celle du nuage des profils-colonnes avec :
   - tableau de données $\mathbf{X} = \mathbf{D}_2^{-1}\mathbf{N}'$ (Le tableau des profils colonnes est $\mathbf{ND}_2^{-1}$ mais, pour garder l'usage de mettre les "individus" en ligne, il faut le transposer ; d'où $\mathbf{X} = \mathbf{D}_2^{-1}\mathbf{N}'$) ;
   - métrique $\mathbf{M} = n\mathbf{D}_1^{-1}$;
   - poids $\mathbf{D} = \dfrac{\mathbf{D}_2}{n}$.

Nous allons voir que leurs résultats sont en dualité exacte.

9.2.1 ACP non centrées et facteur trivial

La matrice de variance d'un nuage de profil est $\mathbf{V} = \mathbf{X}'\mathbf{DX} - \mathbf{gg}'$.

D'après la propriété établie à la fin du paragraphe 9.1.2 $\mathbf{Og}$ est orthogonal au support du nuage, il est donc axe principal, c'est-à-dire vecteur propre de $\mathbf{VM}$, associé à $\lambda = 0$.

Les vecteurs propres de $\mathbf{VM}$ sont alors les mêmes que ceux de $\mathbf{X}'\mathbf{DXM}$ avec les mêmes valeurs propres sauf $\mathbf{g}$ qui a pour valeur propre 1.

En effet $\mathbf{gg}'\mathbf{M}$ est de rang 1 et :

$$\mathbf{X}'\mathbf{DXM} = \mathbf{VM} + \mathbf{gg}'\mathbf{M}$$

d'où :

$$\begin{aligned}\mathbf{X}'\mathbf{DXMg} &= \mathbf{VMg} + \mathbf{gg}'\mathbf{Mg} \\ &= \mathbf{0} + \mathbf{g}\|\mathbf{g}\|^2_{\chi^2} \\ &= \mathbf{g}\end{aligned}$$

Il est donc inutile de centrer les tableaux de profils et on effectuera des ACP non centrées : la valeur propre 1 dont on verra plus tard qu'elle est maximale sera ensuite à éliminer. A cette valeur propre triviale est associé l'axe principal $\mathbf{g}$ et le facteur principal constant :

$$\mathbf{Mg} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \mathbf{1}$$

9.2.2 ACP non centrées des nuages de profils

Il suffit d'appliquer ici les résultats du chapitre 7 : les facteurs principaux sont les vecteurs propres de $\mathbf{MX'DX}$, les composantes principales les vecteurs propres de $\mathbf{XMX'D}$.

Pour les lignes on a $\mathbf{X} = \mathbf{D}_1^{-1}\mathbf{N}$ d'où $\mathbf{X'DX} = \frac{1}{n}\mathbf{N'D}_1^{-1}\mathbf{N}$, $\mathbf{MX'DX} = \mathbf{D}_2^{-1}\mathbf{N'D}_1^{-1}\mathbf{N}$ et $\mathbf{XMX'D} = \mathbf{D}_1^{-1}\mathbf{ND}_2^{-1}\mathbf{N'}$.

Pour les profils-colonnes il suffit de transposer $\mathbf{N}$ et d'inverser les indices 1 et 2, comme le montre le tableau 9.1 :

TABLEAU 9.1

	ACP des profils-lignes	ACP des profils-colonnes
Facteurs principaux	Vecteurs propres de $\mathbf{D}_2^{-1}\mathbf{N'D}_1^{-1}\mathbf{N}$	Vecteurs propres de $\mathbf{D}_1^{-1}\mathbf{ND}_2^{-1}\mathbf{N'}$
Composantes principales	Vecteurs propres de $\mathbf{D}_1^{-1}\mathbf{ND}_2^{-1}\mathbf{N'}$ normalisés par $\mathbf{a'}\frac{\mathbf{D}_1}{n}\mathbf{a} = \lambda$	Vecteurs propres de $\mathbf{D}_2^{-1}\mathbf{N'D}_1^{-1}\mathbf{N}$ normalisés par $\mathbf{b'}\frac{\mathbf{D}_2}{n}\mathbf{b} = \lambda$

On constate que les deux analyses conduisent aux mêmes valeurs propres et que les facteurs principaux de l'une sont les composantes principales de l'autre (à un coefficient multiplicateur près).

En pratique on s'intéresse ici exclusivement aux composantes principales pour obtenir directement les coordonnées factorielles des points représentatifs des profils-lignes ou colonnes. On remarque que les matrices ayant comme vecteurs propres les composantes principales sont les produits des deux matrices $[\mathbf{D}_1^{-1}\mathbf{N}]$ et $[\mathbf{D}_2^{-1}\mathbf{N'}]$ dans un ordre ou dans l'autre :

Les coordonnées des points-lignes et points-colonnes s'obtiennent en cherchant les vecteurs propres des produits des deux tableaux de profils.

La parfaite symétrie entre ACP des profils-lignes et ACP des profils-colonnes conduit alors à superposer les plans principaux des deux ACP afin d'obtenir une représentation simultanée des catégories des deux variables croisées dans le tableau de contingence $\mathbf{N}$. Cette pratique sera justifiée plus en détail au paragraphe 9.4.4

Les cercles de corrélation n'ayant aucun intérêt ici dans le contexte de variables qualitatives l'interprétation des composantes se fait essentiellement en utilisant les contributions des diverses catégories aux inerties des axes factoriels, c'est-à-dire aux valeurs propres.

Comme : $\lambda = \frac{1}{n}\sum_{i=1}^{m_1} n_{i.}(a_i)^2 = \frac{1}{n}\sum_{j=1}^{m_2} n_{.j}(b_j)^2$

On appelle contribution (CTR) du profil-ligne i à l'inertie le quotient :

$$\boxed{\mathrm{CTR}(i) = \frac{\frac{n_{i.}}{n}(a_i)^2}{\lambda}}$$

On a de même : $$\boxed{\mathrm{CTR}(j) = \frac{\frac{n_{.j}}{n}(b_j)^2}{\lambda}}$$

Comme en ACP on considèrera les catégories ayant les plus fortes contributions comme constitutives des axes : un critère simple consistant à retenir les $\mathrm{CTR}(i) > \frac{n_{i.}}{n}$. La contribution doit être complétée par le signe de la coordonnée car certaines catégories peuvent avoir des contributions de sens opposés.

Remarquons qu'ici $\sum_i \frac{n_{i.}}{n} a_i = \sum_j \frac{n_{.j}}{n} b_j = 0$ (les composantes sont centrées) ; il ne peut y avoir d'effet de taille car les coordonnées des points ne peuvent être toutes positives ou toutes négatives.

9.2.3 Formules de transition

Les coordonnées des points-lignes et les coordonnées des points-colonnes sont reliées par des formules simples dont le premier intérêt est d'éviter de réaliser deux diagonalisations. On diagonalisera la matrice la plus petite, par exemple $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$ si $m_1 < m_2$.

Connaissant les solutions $\mathbf{a}$ de l'équation :

$$\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = \lambda\mathbf{a}$$

il suffit de prémultiplier les deux membres de cette équation par $\mathbf{D}_2^{-1}\mathbf{N}'$ pour obtenir un vecteur proportionnel à $\mathbf{b}$:

$$\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = \lambda\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}$$

On a donc $\mathbf{b} = k\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}$. Pour déterminer k il suffit d'utiliser la condition de normalisation $\mathbf{b}'\frac{\mathbf{D}_2}{n}\mathbf{b} = \lambda$ soit $k^2\mathbf{a}'\mathbf{N}\mathbf{D}_2^{-1}\frac{\mathbf{D}_2}{n}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = \frac{k^2}{n}\mathbf{a}'\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = \lambda$. Comme $\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = \lambda\mathbf{D}_1\mathbf{a}$ il vient $\lambda k^2\mathbf{a}'\frac{\mathbf{D}_1}{n}\mathbf{a} = \lambda$ soit $k^2\lambda = 1$ puisque $\mathbf{a}'\frac{\mathbf{D}_1}{n}\mathbf{a} = \lambda$.

On a donc les formules suivantes pour chaque axe :

$$\mathbf{b} = \frac{1}{\sqrt{\lambda}} \mathbf{D}_2^{-1} \mathbf{N'a} \quad \text{soit} \quad b_j = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{.j}} a_i$$

$$\mathbf{a} = \frac{1}{\sqrt{\lambda}} \mathbf{D}_1^{-1} \mathbf{Nb} \quad \text{soit} \quad a_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i.}} b_j$$

avec :

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_{m_1} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_{m_2} \end{bmatrix}$$

Ces formules dites de transition sont des formules pseudo-barycentriques au sens suivant : à $\sqrt{\lambda}$ près la coordonnée d'une catégorie i d'une variable est la moyenne des coordonnées des catégories de l'autre variable pondérées par les fréquences conditionnelles du profil de i.

9.2.4 Trace et reconstitution des données

9.2.4.1 Décomposition du φ^2

Nous avons déjà vu que l'inertie totale des deux nuages était égale au φ^2.

En éliminant la valeur propre triviale on a donc si $m_1 < m_2$:

$$\sum_{k=1}^{m_1 - 1} \lambda_k = \varphi^2$$

car il y a au plus min $((m_1 - 1)\ ;\ (m_2 - 1))$ valeurs propres. Chaque direction principale explique une partie de l'écart à l'indépendance mesurée par le φ^2.

Les pourcentages de variance (ou d'inertie) sont donc les λ_k/φ^2.

P. Cibois (1983) a mis en évidence la propriété suivante qui montre que l'analyse des correspondances étudie la structure des écarts à l'indépendance plus que les écarts eux-mêmes :

Le tableau $\mathbf{N}^*$ défini par :

$$n_{ij}^* = \frac{n_{i.}n_{.j}}{n} + \alpha\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)$$

a mêmes marges que N donc même tableau d'indépendance mais des écarts à l'indépendance réduits de α (si $0 < \alpha < 1$).

L'analyse des correspondances de $\mathbf{N}^*$ est alors presque identique à celle de $\mathbf{N}$: mêmes graphiques, mêmes pourcentages d'inertie, mêmes contributions. Seul φ^2 et les valeurs propres ont changé :

$$(\varphi^2)^* = \alpha^2 \varphi^2 \text{ et } \lambda^* = \alpha^2 \lambda$$

Un utilisateur ne regardant que les pourcentages et non les valeurs absolues ne verrait aucune différence. Le problème est alors de savoir si l'on analyse des écarts significatifs ou non.

9.2.4.2 Formule de reconstitution

La formule $\mathbf{X} = \sum_k \mathbf{c}_k \mathbf{u}'_k \mathbf{M}^{-1}$ établie au chapitre 7 s'applique ici pour $\mathbf{X}$ tableau des profils-lignes, $\mathbf{c}_k$ vecteur des coordonnées des lignes sur l'axe n° k, $\mathbf{u}_k$ facteur principal (identique au vecteur des coordonnées des colonnes sur l'axe k divisé par $\sqrt{\lambda_k}$) et $\mathbf{M} = n\mathbf{D}_2^{-1}$.

On a alors : $$\frac{n_{ij}}{n_{i.}} = \sum_k \frac{a_i^{(k)} b_j^{(k)}}{\sqrt{\lambda_k}} \frac{n_{.j}}{n}$$

mais il faut utiliser tous les facteurs y compris le facteur trivial correspondant à $\lambda = 1$, d'où :

$$\boxed{n_{ij} = \frac{n_{i.} n_{.j}}{n}\left(1 + \sum_k \frac{a_i^{(k)} b_j^{(k)}}{\sqrt{\lambda_k}}\right)}$$

Il s'agit donc bien d'une reconstitution des écarts à l'indépendance à l'aide des coordonnées factorielles des points associés aux profils-lignes et aux profils-colonnes.

9.2.5 Choix du nombre de valeurs propres en AFC

L'AFC est une ACP particulière mais on ne peut appliquer exactement les mêmes règles car la métrique du khi-deux n'est pas la métrique usuelle. On peut ne retenir que les valeurs propres supérieures à leur moyenne comme le fait la règle de Kaiser, mais cette pratique est peu usitée.

La règle du coude reste cependant valide, mais est toujours quelque peu subjective.

Lorsque la taille de l'échantillon le permet, le critère suivant proposé par E. Malinvaud peut se révéler très efficace. Il est basé sur la comparaison entre effectifs observés n_{ij} et effectifs calculés à l'aide de la formule de reconstitution dans le contexte suivant : on fait l'hypothèse que les données forment un échantillon tiré aléatoirement et avec équiprobabilité dans une population telle que $p_{ij} = p_{i.} p_{.j}\left(1 + \sum_{k=1}^{K} \alpha_{ik} \beta_{jk}\right)$. En d'autres termes la loi bidimensionnelle sous-jacente est un tableau de rang K.

Dans ces conditions, si $\hat{n}_{ij}^{(K)} = \left(\frac{n_{i.} n_{.j}}{n}\right)\left(1 + \sum_{k=1}^{K} a_{ik} b_{jk} \Big/ \sqrt{\lambda_k}\right)$ est la reconstitution de la case ij, à l'aide des K premiers axes, on peut montrer que la quantité

$$Q_K = \sum_{i,j} \frac{(n_{ij} - \hat{n}_{ij}^{K})^2}{\frac{n_{i.} n_{.j}}{n}}$$

suit asymptotiquement une loi du χ^2 à $(p - K - 1)$ $(q - K - 1)$ degrés de liberté. Il s'agit donc d'une généralisation du test d'écart à l'indépendance qui correspond au cas $K = 0$.

On trouve sans difficulté que Q_K se calcule à l'aide des valeurs propres et est égal à n fois l'inertie au delà de la dimension K :

$$Q_K = n(I - \lambda_1 - \lambda_2 - \cdots - \lambda_K) = n(\lambda_{K+1} + \lambda_{K+2} + \cdots + \lambda_r)$$

où $r = min(p - 1 \; ; q - 1)$

On peut donc tester successivement les valeurs de K depuis $K = 0$ (hypothèse d'indépendance), jusqu'au moment où on ne peut plus rejeter l'ajustement.

Les conditions d'application sont celles du test du khi-deux : effectifs théoriques au moins égaux à 5. Cependant si n est très élevé le test conduit à conserver un trop grand nombre de valeurs propres : on ne l'emploiera que pour n inférieur à quelques milliers.

9.3 UN EXEMPLE

Nous avons soumis à l'analyse des correspondances (logiciel SPAD Version 5.6) le tableau de contingence sur les vacances des français en 1999 déjà étudié dans le chapitre 6.

Le tableau des valeurs propres montre clairement que deux axes suffisent à décrire la liaison entre la catégorie socio-professionnelle et le mode d'hébergement :

SOMME DES VALEURS PROPRES . . . 0.1073
HISTOGRAMME DES 7 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	0.0657	61.24	61.24	**********************************
2	0.0254	23.70	84.94	*******************
3	0.0081	7.55	92.49	**********
4	0.0037	3.46	95.95	*****
5	0.0028	2.60	98.55	****
6	0.0014	1.29	99.84	**
7	0.0002	0.16	100.00	*

Ici le test de Malinvaud est inopérant car $n = 18352$ est trop élevé.

Les tableaux suivants permettent de repérer les modalités ayant des contributions significatives : Sur l'axe 1 *Hotel, Résidence secondaire* liés avec *retraités* et opposés à *tente* et *ouvrier*. L'axe 2 est caractérisé par *Résidence secondaire de parents et amis* et *cadres*.

On retrouve des associations détectées par la décomposition (figure 9.3) du khi-deux, mais le graphique permet de les illustrer de manière évocatrice.

Rappelons que l'interprétation des proximités sur le graphique doit respecter certains principes : si deux modalités d'une même variable sont proches et bien représentées, cela signifie que leurs profils sont semblables (c'est le cas d'*ouvriers* et *employés* par exemple qui fréquentent les mêmes lieux dans des proportions proches). Par contre la proximité entre une modalité d'une variable et une modalité de l'autre, comme *profession intermédiaire* et *village de vacances*, est plus délicate à interpréter : ce que l'on peut seulement dire c'est que le barycentre des 3787 séjours des *professions intermédiaires* est proche du barycentre des 686 séjours en *village de vacances* (voir plus loin).

MODES D'HEBERGEMENT			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
HOTE - Hotel	14.19	0.11	-0.32	-0.03	0.05	-0.03	-0.08	21.6	0.4	3.9	4.3	29.1	0.90	0.01	0.02	0.01	0.05
LOCA - Location	16.68	0.04	0.13	-0.11	0.12	0.03	-0.01	4.1	7.6	28.2	3.3	1.3	0.38	0.27	0.32	0.02	0.01
RSEC - Résid.	9.36	0.33	-0.55	0.15	0.00	-0.01	0.07	43.2	8.4	0.0	0.6	15.2	0.91	0.07	0.00	0.00	0.01
RPPA - Résid. Par	33.73	0.03	0.11	0.09	-0.10	0.02	-0.02	6.3	10.7	38.4	3.9	4.8	0.40	0.27	0.30	0.01	0.01
RSPA - Résid. Sec.	9.98	0.15	-0.06	-0.37	-0.07	0.04	0.07	0.5	53.9	6.8	3.9	17.0	0.02	0.90	0.04	0.01	0.03
TENT - Tente	4.17	0.31	0.52	-0.02	0.01	-0.19	0.01	17.1	0.1	0.0	40.2	0.3	0.86	0.00	0.00	0.11	0.00
CARA - Caravane	6.10	0.18	0.25	0.27	0.16	0.09	0.08	6.0	17.3	19.6	12.2	13.2	0.36	0.41	0.15	0.04	0.03
AJ - Auberge	2.09	0.07	0.15	-0.13	0.10	0.00	-0.10	0.7	1.4	2.7	0.0	7.3	0.33	0.27	0.16	0.00	0.15
VILL - Village	3.70	0.07	0.10	0.03	0.02	-0.18	0.09	0.5	0.2	0.2	31.7	11.7	0.14	0.02	0.01	0.49	0.14

CATEGORIES SOCIO-PROFESSIONNELLES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Agriculteurs	1.27	0.27	0.15	-0.14	0.13	-0.44	-0.11	0.4	0.9	2.6	67.3	5.5	0.08	0.07	0.06	0.73	0.04
Artisans,	6.15	0.08	-0.01	-0.15	0.18	0.03	-0.14	0.0	5.6	24.6	1.2	44.6	0.00	0.30	0.42	0.01	0.26
Cadres,	23.47	0.07	-0.13	-0.23	-0.03	0.02	0.02	5.6	49.9	3.0	2.5	3.6	0.22	0.75	0.01	0.01	0.01
Prof. interm.	20.43	0.01	0.08	-0.01	0.02	-0.04	0.06	2.0	0.0	0.9	7.6	23.6	0.46	0.00	0.02	0.10	0.23
Employés	10.57	0.08	0.26	0.06	-0.04	0.08	-0.04	10.9	1.3	1.7	16.3	6.5	0.82	0.04	0.02	0.07	0.02
Ouvriers	16.56	0.14	0.33	0.13	0.05	0.01	0.02	28.2	11.8	4.2	0.3	2.1	0.83	0.13	0.02	0.00	0.00
Retraités	19.44	0.20	-0.41	0.20	-0.01	0.00	-0.01	48.6	30.3	0.3	0.0	0.9	0.80	0.19	0.00	0.00	0.00
Autres inactifs	2.12	0.40	0.36	0.01	-0.49	-0.09	-0.13	4.2	0.0	62.8	4.7	13.3	0.33	0.00	0.60	0.02	0.04

Dans la figure 9.3, les points ont des tailles proportionnelles à leurs fréqences marginales.

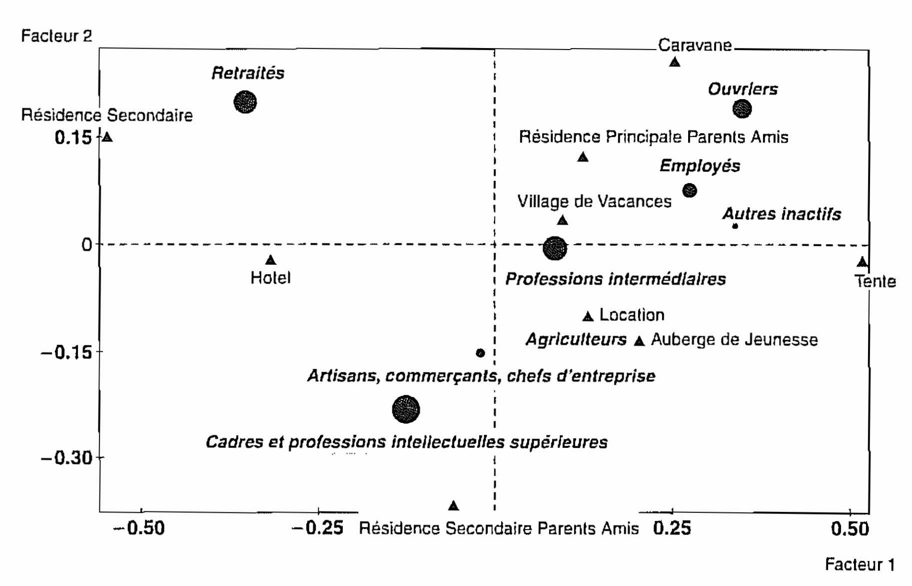


FIGURE 9.3

9.4 ANALYSE CANONIQUE DE DEUX VARIABLES QUALITATIVES, JUSTIFICATION DE LA REPRÉSENTATION SIMULTANÉE

9.4.1 Mise sous forme disjonctive de données qualitatives

Le tableau de contingence N ne constitue pas en réalité le tableau de données brutes : il est le résultat d'un traitement élémentaire (tri croisé) de données relevées sur n individus du type : $(x_i^1 ; x_i^2)$ pour $i = 1, 2, \ldots, n$ où x_i^1 et x_i^2 sont les numéros des catégories des variables qualitatives $\mathscr{X}_1$ et $\mathscr{X}_2$. La numérotation des catégories est arbitraire et on introduit alors la représentation suivante comme au paragraphe 6.4.3 :

– A une variable qualitative $\mathscr{X}$ à m catégories on associe les m variables indicatrices de ses catégories : $\mathbb{1}^1, \mathbb{1}^2, \ldots \mathbb{1}^m$. $\mathbb{1}^x(i)$ vaut 1 si x est dans la catégorie i, 0 sinon. Pour un individu i une seule des m indicatrices vaut 1 les $m - 1$ autres valent 0.

Pour n individus la variable $\mathscr{X}$ peut être représentée par le tableau de données binaires **X** suivant :

$$\mathbf{X} = \begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & \ldots & & m \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ & & & & \\ & & & & \\ & & & & \\ 0 & 0 & 1 & \ldots & 0 \end{bmatrix} \end{array}$$

On dit que $\mathcal{X}$ a été mise sous forme disjonctive.

On remarque que $\sum_{x=1}^{m} \mathbb{1}^x = 1$ donc que les m colonnes de $\mathbf{X}$ ont pour somme le vecteur $\mathbf{1}$.

– A deux variables qualitatives $\mathcal{X}_1$ et $\mathcal{X}_2$ correspondent donc deux matrices $\mathbf{X}_1$ et $\mathbf{X}_2$ à n lignes et respectivement m_1 et m_2 colonnes.

On vérifie alors les formules suivantes liant $\mathbf{X}_1$ et $\mathbf{X}_2$ au tableau de contingence $\mathbf{N}$ et aux deux matrices d'effectifs marginaux $\mathbf{D}_1$ et $\mathbf{D}_2$:

$$\mathbf{N} = \mathbf{X}_1' \mathbf{X}_2$$
$$\mathbf{D}_1 = \mathbf{X}_1' \mathbf{X}_1$$
$$\mathbf{D}_2 = \mathbf{X}_2' \mathbf{X}_2$$

En effet, faire le produit scalaire de deux vecteurs d'indicatrices revient à compter le nombre de co-occurrences.

9.4.2 Quantifications de variables qualitatives

Si à chaque catégorie d'une variable qualitative $\mathcal{X}$ on associe une valeur numérique, on transforme $\mathcal{X}$ en une variable discrète à m valeurs : on réalise ainsi une quantification de $\mathcal{X}$ en une variable numérique x (certains auteurs parlent de « codage »). Il existe une infinité de quantifications possibles dont la structure est celle d'un sous-espace vectoriel de l'espace des variables.

Si a_j est la valeur numérique associée à la catégorie j, on a :

$$x = \sum_{j=1}^{m} a_j \mathbb{1}^j$$

Une quantification n'est donc qu'une combinaison linéaire des variables indicatrices.
Pour l'ensemble des n individus on a :

$$x_i = \sum_{j=1}^{m} a_j \mathbb{1}^j(i) \qquad \text{soit si } \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_m \end{bmatrix}$$

$$\mathbf{x} = \mathbf{X}\mathbf{a}$$

L'ensemble des $\mathbf{x}$ est donc le sous-espace W engendré par les combinaisons linéaires des colonnes de $\mathbf{X}$.

9.4.3 Analyse canonique de deux groupes d'indicatrices

L'étude de la dépendance entre $\mathcal{X}_1$ et $\mathcal{X}_2$ est donc celle des relations entre les deux groupes de variables indicatrices associées. On peut donc appliquer l'analyse canonique étudiée au chapitre précédent.

Les deux tableaux de données à analyser sont les tableaux disjonctifs $\mathbf{X}_1$ et $\mathbf{X}_2$. On constate immédiatement que les deux espaces W_1 et W_2 engendrés par les colonnes de ces tableaux ont en commun le vecteur $\mathbf{1}$ qui est le vecteur somme des colonnes de $\mathbf{X}_1$ ou de $\mathbf{X}_2$ (donc $\dim(W_1 \cap W_2) \geq 1$). Les variables canoniques autres que $\mathbf{1}$ formant des systèmes $\mathbf{D}$-orthonormés de W_1 et W_2, sont donc centrées, car elles sont orthogonales au vecteur $\mathbf{1}$.

En supposant ici que les n individus ont mêmes poids $1/n$, avec les notations du chapitre 8 on a :

$$\mathbf{V}_{11} = \frac{1}{n}\mathbf{X}_1'\mathbf{X}_1 = \frac{1}{n}\mathbf{D}_1$$

$$\mathbf{V}_{22} = \frac{1}{n}\mathbf{X}_2'\mathbf{X}_2 = \frac{1}{n}\mathbf{D}_2$$

$$\mathbf{V}_{12} = \frac{1}{n}\mathbf{X}_1'\mathbf{X}_2 = \frac{1}{n}\mathbf{N}$$

$$\mathbf{V}_{21} = \frac{1}{n}\mathbf{N}'$$

Les facteurs canoniques du groupe 1 sont les vecteurs propres de $\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$ c'est-à-dire de $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$: ce sont donc les composantes principales de l'ACP des profils-lignes à un facteur multiplicatif près.

De même les facteurs canoniques du groupe 2 sont les vecteurs propres de $\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}$ et fournissent les coordonnées des profils-colonnes sur les axes principaux. Les valeurs propres λ de l'analyse des correspondances sont donc les carrés des coefficients de corrélation canonique (ce qui prouve que $\lambda \leq 1$) et la valeur propre triviale $\lambda = 1$ correspond au fait que W_1 et W_2 ont $\mathbf{1}$ dans leur intersection.

Les facteurs canoniques donnent des combinaisons linéaires des variables indicatrices, donc des quantifications de $\mathcal{X}_1$ et $\mathcal{X}_2$: on peut interpréter l'analyse des correspondances comme la recherche d'une transformation simultanée de $\mathcal{X}_1$ et $\mathcal{X}_2$ en variables numériques telles que leur coefficient de corrélation linéaire soit maximal. Cette présentation plus connue des statisticiens anglophones est attribuée à Fisher, elle remonte en fait à des travaux de Hirschfeld, alias H.O. Hartley, de 1936.

Les valeurs numériques optimales à attribuer aux catégories sont donc leurs coordonnées sur le premier axe de l'analyse des correspondances. Si l'on réordonne lignes et colonnes du tableau de contingence $\mathbf{N}$ selon l'ordre des points sur le premier axe principal on obtient un tableau tel que les termes « diagonaux » aient des effectifs maximaux.

Les formules de transition sont identiques à celles permettant de passer des facteurs canoniques d'un groupe à ceux de l'autre groupe.

Il n'est donc pas nécessaire dans ce contexte d'introduire la métrique du χ^2 et on voit que les catégories des deux variables $\mathcal{X}_1$ et $\mathcal{X}_2$ sont traitées de la même façon en tant qu'éléments de $\mathbb{R}^n$ grâce aux variables indicatrices ce qui justifie le fait de les représenter simultanément sur les mêmes graphiques.

Les représentations graphiques de l'analyse canonique (cercle des corrélations) sont cependant ici inadéquates car la notion de corrélation avec une variable indicatrice n'a guère de sens : on se contentera de représenter chaque catégorie par ses « codages » successifs sur les axes.

9.4.4 Représentation simultanée optimale des $(m_1 + m_2)$ catégories d'individus

Les catégories des variables qualitatives $\mathcal{X}_1$ et $\mathcal{X}_2$ définissent des sous-groupes d'individus d'effectifs $n_{i.}$ $(i = 1, 2, \ldots, m_1)$ et $n_{.j}$ $(j = 1, 2, \ldots, m_2)$. Si l'on dispose d'une variable numérique $\mathbf{z}$ de moyenne nulle représentant les coordonnées des n individus sur un axe on représentera la catégorie i de $\mathcal{X}_1$ par un point dont la coordonnée a_i est la moyenne des coordonnées des $n_{i.}$ individus de la catégorie en question :

$$a_i = \frac{1}{n_{i.}} \sum_{k=1}^{n} z_k \mathbb{1}^i(k) = \frac{1}{n_{i.}} \mathbf{z}'\mathbf{x}_1^i$$

où $\mathbf{x}_1^i$ est la $i^{\text{ème}}$ colonne de $\mathbf{x}_1$.

On en déduit que le vecteur $\mathbf{a}$ renfermant les coordonnées des m_1 catégories de $\mathbf{X}_1$ est :

$$\mathbf{a} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{z} = \mathbf{D}_1^{-1}\mathbf{X}_1'\mathbf{z}$$

de même pour les m_2 catégories de $\mathbf{X}_2$:

$$\mathbf{b} = \mathbf{D}_2^{-1}\mathbf{X}_2'\mathbf{z}$$

La variable $\mathbf{z}$ est d'autant plus intéressante pour $\mathbf{X}_1$ qu'elle permet de bien séparer les a_i, c'est-à-dire que la variance $\frac{1}{n}\,\mathbf{a}'\mathbf{D}_1\mathbf{a}$ est plus grande. Le maximum de cette variance est obtenu si tous les individus appartenant à une même catégorie de $\mathcal{X}_1$ ont la même valeur de $\mathbf{z}$.

Cherchons alors la variable $\mathbf{z}$ et les coordonnées $\mathbf{a}$ et $\mathbf{b}$ telles que en moyenne $\mathbf{a}'\mathbf{D}_1\mathbf{a}$ et $\mathbf{b}'\mathbf{D}_2\mathbf{b}$ soient maximales : on aura alors en un certain sens une représentation simultanée optimale des catégories des deux variables sur un axe.

Comme $\mathbf{a}'\mathbf{D}_1\mathbf{a} = \mathbf{z}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{z} = \mathbf{z}'\mathbf{A}_1\mathbf{z}$ où $\mathbf{A}_1$ est le projecteur sur W_1 ; et $\mathbf{b}'\mathbf{D}_2\mathbf{b} = \mathbf{z}'\mathbf{A}_2\mathbf{z}$, le maximum de $\frac{1}{2}\,[\mathbf{a}'\mathbf{D}_1\mathbf{a} + \mathbf{b}'\mathbf{D}_2\mathbf{b}]$ s'obtient lorsque $\frac{1}{2}\,[\mathbf{z}'(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{z}]$ est maximal. En supposant $V(\mathbf{z})$ fixé ce maximum est atteint pour $\mathbf{z}$ vecteur propre de $\frac{1}{2}(\mathbf{A}_1 + \mathbf{A}_2)$:

$$(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{z} = 2\mu\mathbf{z}$$

Comme $\mathbf{A}_1\mathbf{z} = \mathbf{X}_1\mathbf{D}_1^{-1}\mathbf{X}_1'\mathbf{z} = \mathbf{X}_1\mathbf{a}$ et $\mathbf{A}_2\mathbf{z} = \mathbf{X}_2\mathbf{D}_2^{-1}\mathbf{X}_2'\mathbf{z} = \mathbf{X}_2\mathbf{b}$ il vient :

$$\mathbf{X}_1\mathbf{a} + \mathbf{X}_2\mathbf{b} = 2\mu\mathbf{z}$$

soit en prémultipliant les deux membres de cette équation par $\mathbf{D}_1^{-1}\mathbf{X}_1'$:

$$\mathbf{a} + \mathbf{D}_1^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b} = 2\mu\mathbf{a}$$

soit :

$$\mathbf{a} + \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = 2\mu\mathbf{a}$$

et en prémultipliant par $\mathbf{D}_2^{-1}\mathbf{X}_2'$:

$$\mathbf{D}_2^{-1}\mathbf{X}_2'\mathbf{X}_1\mathbf{a} + \mathbf{b} = 2\mu\mathbf{b}$$

ou

$$\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} + \mathbf{b} = 2\mu\mathbf{b}$$

il vient alors :

$$\begin{cases} \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = (2\mu - 1)\mathbf{a} \\ \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = (2\mu - 1)\mathbf{b} \end{cases}$$

On reconnaît les formules de transition et par substitution on a :

$$\begin{cases} \mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = (2\mu - 1)^2\mathbf{a} \\ \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = (2\mu - 1)^2\mathbf{b} \end{cases}$$

Ce sont les équations de l'analyse factorielle des correspondances avec $(2\mu - 1)^2 = \lambda$.

Remarquons que l'on aurait pu appliquer directement les résultats du paragraphe 8.3.1 du chapitre précédent : $\mathbf{z}$ est alors le compromis à un facteur près des deux variables canoniques $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$.

Les coordonnées des points catégories données par le premier axe de l'analyse des correspondances sont donc optimales ; les axes suivants correspondent au choix d'autres variables $\mathbf{z}$ orthogonales aux précédentes.

La signification réelle de la représentation simultanée est donc celle-ci : ***les points représentatifs des catégories des deux variables sont les barycentres des groupes d'individus qu'elles définissent.***

Les proximités entre points représentatifs doivent être interprétées comme des proximités entre moyennes : pour deux catégories i et i' d'une même variable cela entraîne une proximité de leurs profils. Pour deux catégories i et j l'une de $\mathscr{X}_1$ l'autre de $\mathscr{X}_2$ l'interprétation est plus délicate.

On peut également représenter sur le graphique les cases du tableau de contingence : tout individu de la catégorie i de $\mathscr{X}_1$ et j de $\mathscr{X}_2$ a pour coordonnée z sur un axe :

$$z = \frac{1}{2\mu}(a_i + b_j) = \frac{1}{1 + \sqrt{\lambda}}(a_i + b_j)$$

d'après la formule $\mathbf{X}_1\mathbf{a} + \mathbf{X}_2\mathbf{b} = 2\mu\mathbf{z}$.

9.4.5 La méthode des moyennes réciproques

La présentation suivante connue sous le nom de *reciprocal averaging* ou de *dual scaling* éclaire également la représentation simultanée de l'analyse des correspondances.

Supposons que l'on place sur un axe les catégories de la variable $\mathcal{X}_1$ comme des points de coordonnées a_i. Par exemple les CSP dans le cas étudié précédemment :

CSP 1 CSP 2 CSP m_1

0

Pour représenter une catégorie j de l'autre variable $\mathcal{X}_2$, ici le mode d'hébergement, il semble logique de la représenter comme le barycentre des catégories professionnelles avec pour coefficients les importances relatives des diverses CSP dans le mode d'hébergement en question :

$$b_j = \sum_{j=1}^{m_1} \frac{n_{ij}}{n_{.j}} a_i \qquad \text{soit } \mathbf{b} = \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}$$

L'idéal serait que la réciproque soit vraie, c'est-à-dire que l'on puisse représenter les catégories de $\mathcal{X}_1$ comme barycentres des catégories de $\mathcal{X}_2$:

$$\mathbf{a} = \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b}$$

La simultanéité de ces deux relations est impossible : on cherchera alors une représentation barycentrique simultanée approchée avec :

$$\begin{cases} \alpha\mathbf{a} = \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} \\ \alpha\mathbf{b} = \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} \end{cases}$$

où α est le plus grand possible car $\alpha < 1$. On retrouve alors les équations de l'analyse des correspondances avec $\alpha = \sqrt{\lambda}$.

L'algorithme consistant à partir d'un vecteur $\mathbf{a}^0$ arbitraire, à en déduire $\mathbf{b}^{(1)} = \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}^0$ puis $\mathbf{a}^{(1)} = \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b}^{(1)}$, etc., avec normalisation à chaque étape jusqu'à convergence fournit en général la première solution de l'analyse des correspondances relative à λ_1.

9.4.6 Conclusion

L'analyse des correspondances est la méthode privilégiée d'étude des relations entre deux variables qualitatives et l'une de ses principales propriétés est la faculté de représenter simultanément lignes et colonnes d'un tableau de contingence. Si en théorie elle ne s'applique qu'à des tableaux de contingence, elle peut être étendue moyennant certaines précautions à d'autres types de tableaux comme le prouvera le chapitre suivant.

10 L'analyse des correspondances multiples

L'analyse des correspondances multiples (ACM) est une technique de description de données qualitatives : on considère ici n individus décrits par p variables $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_p$ à $m_1, m_2, \ldots, m_p$ catégories. Cette méthode est particulièrement bien adaptée à l'exploration d'enquêtes où les questions sont à réponses multiples.

Sur le plan formel il s'agit d'une simple application de l'analyse des correspondances au tableau disjonctif des $m_1 + m_2 + \cdots + m_p$ indicatrices des catégories. Cette méthode possède cependant des propriétés qui la relient à d'autres méthodes statistiques et lui donnent son statut particulier et en font l'équivalent de l'analyse en composantes principales pour des variables qualitatives.

10.1 PRÉSENTATION FORMELLE

10.1.1 Données et notations

Chaque individu est décrit par les numéros des catégories des p variables auxquelles il appartient. Ces données brutes se présentent donc sous forme d'un tableau à n lignes et p colonnes. Les éléments de ce tableau sont des codes arbitraires sur lesquels aucune opération arithmétique n'est licite. La forme mathématique utile pour les calculs est alors le tableau disjonctif des indicatrices des p variables obtenu en juxtaposant les p tableaux d'indicatrices de chaque variable $\mathscr{X}_i$.

Ainsi le tableau brut suivant :

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 2 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

correspondant à 5 observations de trois variables $\mathscr{X}_1$, $\mathscr{X}_2$, $\mathscr{X}_3$ à 3, 2, 3 catégories respectivement engendre le tableau disjonctif X à 5 lignes et 8 colonnes :

$$\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2|\mathbf{X}_3) = \begin{bmatrix} 1 & 0 & 0 & \vdots & 0 & 1 & \vdots & 0 & 0 & 1 \\ 0 & 1 & 0 & \vdots & 1 & 0 & \vdots & 1 & 0 & 0 \\ 0 & 1 & 0 & \vdots & 0 & 1 & \vdots & 0 & 1 & 0 \\ 0 & 0 & 1 & \vdots & 0 & 1 & \vdots & 1 & 0 & 0 \\ 0 & 0 & 1 & \vdots & 1 & 0 & \vdots & 0 & 1 & 0 \end{bmatrix}$$

La somme des éléments de chaque ligne de $\mathbf{X}$ est égale à p, nombre de variables.

La somme des éléments d'une colonne de $\mathbf{X}$ donne l'effectif marginal de la catégorie correspondante.

La somme des colonnes de chaque tableau d'indicatrices est égale au vecteur $\mathbf{1}$; le rang de $\mathbf{X}$ est donc $\sum_{i=1}^{p} m_i - p + 1$.

On notera $\mathbf{D}$ le tableau diagonal des effectifs marginaux des $m_1 + m_2 + \cdots + m_p$ catégories :

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & & & \mathbf{0} \\ & \mathbf{D}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{D}_p \end{bmatrix}$$

10.1.2 Une propriété remarquable pour $p = 2$

Pour deux variables qualitatives $\mathscr{X}_1$ et $\mathscr{X}_2$ à m_1 et m_2 modalités, l'analyse factorielle des correspondances du tableau disjonctif $X = (X_1|X_2)$ est équivalente à l'analyse factorielle des correspondances (AFC) du tableau de contingence $N = X_1'X_2$.

Cette propriété est à l'origine du nom de la méthode étudiée ici.

10.1.2.1 AFC formelle du tableau disjonctif

L'AFC d'un tableau $\mathbf{X}$ revient à chercher les valeurs propres et les vecteurs propres du produit des deux tableaux de profils associés à $\mathbf{X}$.

Le tableau des profils-lignes vaut ici $\mathbf{X}/2$.

Le tableau des profils des colonnes $\mathbf{XD}^{-1}$ est tel que :

$$\mathbf{XD}^{-1} = (\mathbf{X}_1|\mathbf{X}_2)\begin{bmatrix} \mathbf{D}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2^{-1} \end{bmatrix}$$

Les coordonnées des profils des colonnes sont les vecteurs propres de :

$$(\mathbf{X}\mathbf{D}^{-1})'\frac{1}{2}\mathbf{X} = \frac{1}{2}\mathbf{D}^{-1}\mathbf{X}'\mathbf{X}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix}\mathbf{X}_1' \\ \hline \mathbf{X}_2'\end{bmatrix}[\mathbf{X}_1|\mathbf{X}_2] = \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix} = \begin{bmatrix}\mathbf{D}_1 & \mathbf{N} \\ \mathbf{N}' & \mathbf{D}_2\end{bmatrix}$$

L'équation donnant les $m_1 + m_2$ coordonnées des profils des colonnes est, en notant **a** les m_1 premières composantes et **b** les m_2 suivantes :

$$\frac{1}{2}\begin{bmatrix}\mathbf{D}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{D}_1 & \mathbf{N} \\ \mathbf{N}' & \mathbf{D}_2\end{bmatrix}\begin{bmatrix}\mathbf{a} \\ \mathbf{b}\end{bmatrix} = \mu\begin{bmatrix}\mathbf{a} \\ \mathbf{b}\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{I}_{m_1} & \mathbf{D}_1^{-1}\mathbf{N} \\ \mathbf{D}_2^{-1}\mathbf{N}' & \mathbf{I}_{m_2}\end{bmatrix}\begin{bmatrix}\mathbf{a} \\ \mathbf{b}\end{bmatrix} = 2\mu\begin{bmatrix}\mathbf{a} \\ \mathbf{b}\end{bmatrix}$$

d'où les équations :

$$\begin{cases}\mathbf{a} + \mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = 2\mu\mathbf{a} \\ \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} + \mathbf{b} = 2\mu\mathbf{b}\end{cases} \quad \text{ou} \quad \begin{cases}\mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = (2\mu - 1)\mathbf{a} \\ \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = (2\mu - 1)\mathbf{b}\end{cases}$$

On reconnaît les équations de l'analyse des correspondances de **N** (formules de transition) et par substitution :

$$\begin{cases}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = (2\mu - 1)^2\mathbf{b} \\ \mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = (2\mu - 1)^2\mathbf{a}\end{cases}$$

avec $\lambda = (2\mu - 1)^2$.

Les coordonnées des $m_1 + m_2$ points colonnes de **X** sont donc identiques (à un coefficient de proportionnalité près) aux coordonnées des lignes et des colonnes de **N** dans la représentation simultanée.

10.1.2.2 Propriétés particulières des valeurs propres et vecteurs propres

Si $n > m_1 + m_2$, l'AFC du tableau **X** va aboutir à plus de facteurs que l'AFC de N.

D'où viennent les solutions supplémentaires? Notons tout d'abord l'existence d'une solution triviale supplémentaire correspondant à une valeur propre nulle puisque les colonnes de **X** sont liées par une relation linéaire (la somme des colonnes de $\mathbf{X}_1$ est égale à la somme des colonnes de $\mathbf{X}_2$). Il y a donc $m_1 + m_2 - 2$ valeurs propres non trivialement nulles ou égales à 1.

Comme $\lambda = (2\mu - 1)^2$, à chaque λ correspondent deux valeurs propres :

$$\mu = \frac{1 + \sqrt{\lambda}}{2} \quad \text{et} \quad \mu = \frac{1 - \sqrt{\lambda}}{2}$$

correspondant aux vecteurs propres $\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}$ et $\begin{pmatrix} \mathbf{a} \\ -\mathbf{b} \end{pmatrix}$ soit, si $m_1 < m_2$, $2(m_1 - 1)$ valeurs propres.

Il y a en plus $m_2 - m_1$ vecteurs propres du type $\begin{pmatrix} \mathbf{0} \\ \mathbf{b} \end{pmatrix}$ associés à la valeur propre 1/2 qui est donc de multiplicité $m_2 - m_1$.

Seules les $(m_1 - 1)$ valeurs propres supérieures à 1/2 ont une signification.

Comme :
$$\text{Trace}\begin{bmatrix} \mathbf{I}_{m_1} & \mathbf{D}_1^{-1}\mathbf{N} \\ \mathbf{D}_2^{-1}\mathbf{N}' & \mathbf{I}_{m_2} \end{bmatrix} = m_1 + m_2$$

l'inertie totale est égale à $\dfrac{m_1 + m_2}{2} - 1$.

Bien que fournissant des axes identiques à l'analyse des correspondances de **N**, les inerties associées et les parts d'inertie sont très différentes et ne peuvent être interprétées sans précaution.

Ainsi l'analyse des correspondances sur le tableau disjonctif associé au tableau étudié au chapitre précédent conduit aux résultats suivants : ($m_1 = 9$ et $m_2 = 8$) :

$\mu_1 = 0.628$ 8.37%	$\lambda_1 = 0.0657$ 61.24%
$\mu_2 = 0.580$ 7.77%	$\lambda_2 = 0.0254$ 23.7%
$\mu_3 = 0.545$ 7.27%	$\lambda_3 = 0.0081$ 7.55%
$\sum_{i=1}^{15} \mu_i = 7.5 = \dfrac{m_1 + m_2}{2} - 1$	$\sum_{i=1}^{7} \lambda_i = 0.1073$

Les valeurs propres qui étaient très séparées dans l'AFC de **N**, ne le sont plus dans l'AFC de **X**.

10.1.3 Le cas général $p > 2$

La propriété précédente conduit à l'extension à p variables qui consiste à effectuer une analyse des correspondances sur le tableau disjonctif $\mathbf{X} = (\mathbf{X}_1 | \mathbf{X}_2 | \ldots \mathbf{X}_p)$ pour obtenir ainsi une représentation simultanée des $m_1 + m_2 + \cdots + m_p$ catégories comme points d'un espace de faible dimension.

10.1.3.1 Coordonnées des catégories

On notera $\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{a}_p \end{bmatrix}$ le vecteur à $\sum_{i=1}^{p} m_i$ composantes des coordonnées factorielles des catégories de toutes les variables sur un axe.

Pour chaque valeur propre μ on a donc :

$$\frac{1}{p}\mathbf{D}^{-1}\mathbf{X}'\mathbf{X}\mathbf{a} = \mu\mathbf{a}$$

soit :

$$\frac{1}{p}\begin{bmatrix} \mathbf{D}_1^{-1} & & & \mathbf{0} \\ & \mathbf{D}_2^{-1} & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{D}_p^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \dots & \mathbf{X}_1'\mathbf{X}_p \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \dots & \\ \vdots & & & \\ \mathbf{X}_p'\mathbf{X}_1 & \dots & & \mathbf{X}_p'\mathbf{X}_p \end{bmatrix}\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_p \end{bmatrix} = \mu\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_p \end{bmatrix}$$

On note $\mathbf{B}$ le tableau dont les blocs sont les $\mathbf{X}_i'\mathbf{X}_j$. Ce tableau, dit tableau de Burt, est un super-tableau de contingence puisque chaque bloc $\mathbf{X}_i'\mathbf{X}_j$ est le tableau de contingence croisant $\mathcal{X}_i$ avec $\mathcal{X}_j$.

L'équation des coordonnées des catégories est donc :

$$\boxed{\frac{1}{p}\mathbf{D}^{-1}\mathbf{B}\mathbf{a} = \mu\mathbf{a}}$$

On prendra comme convention de normalisation :

$$\boxed{\frac{1}{np}\mathbf{a}'\mathbf{D}\mathbf{a} = \mu}$$

car la somme des éléments de $\mathbf{X}$ vaut np.

10.1.3.2 Coordonnées des individus

Les lignes de $\mathbf{X}$ représentant les individus, les coordonnées des points-lignes s'obtiennent en diagonalisant le produit, effectué dans l'ordre inverse, des deux tableaux des profils. Soit $\mathbf{z}$ le vecteur à n composantes des coordonnées des n individus sur un axe factoriel. On a :

$$\boxed{\frac{1}{p}\mathbf{X}\mathbf{D}^{-1}\mathbf{X}'\mathbf{z} = \mu\mathbf{z}}$$

En développant par blocs $\mathbf{X}\mathbf{D}^{-1}\mathbf{X}'$ il vient :

$$\frac{1}{p}(\mathbf{X}_1|\mathbf{X}_2|\dots|\mathbf{X}_p)\begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & & & \mathbf{0} \\ & (\mathbf{X}_2'\mathbf{X}_2)^{-1} & & \\ & & \ddots & \\ \mathbf{0} & & & (\mathbf{X}_p'\mathbf{X}_p)^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{X}_1' \\ \hline \mathbf{X}_2' \\ \hline \vdots \\ \hline \mathbf{X}_p' \end{bmatrix}\mathbf{z} = \mu\mathbf{z}$$

soit $\frac{1}{p}\left(\sum_{i=1}^{p} \mathbf{X}_i(\mathbf{X}_i' \mathbf{X}_i)^{-1}\mathbf{X}_i'\right)\mathbf{z} = \mu\mathbf{z} = \frac{1}{p}\sum_{i=1}^{p}\mathbf{A}_i\mathbf{z}$.

$\mathbf{X}_i(\mathbf{X}_i' \mathbf{X}_i)^{-1}\mathbf{X}_i'$ est le projecteur orthogonal $\mathbf{A}_i$ sur l'espace engendré par les combinaisons linéaires des indicatrices des catégories de $\mathcal{X}_i$.

$\mathbf{z}^0 = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}$ étant la solution triviale associée à $\mu = 1$ les autres solutions lui sont orthogonales.

Les coordonnées des individus sur un axe sont donc de moyenne nulle.

La condition habituelle de normalisation est :

$$\frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\mathbf{z}'\mathbf{z} = \mu$$

10.1.3.3 Formules de transition et relations barycentriques

D'après les résultats du chapitre précédent on a :

$$\boxed{\mathbf{z} = \frac{1}{\sqrt{\mu}}\frac{1}{p}\mathbf{Xa}} \quad \text{et} \quad \boxed{\mathbf{a} = \frac{1}{\sqrt{\mu}}\mathbf{D}^{-1}\mathbf{X}'\mathbf{z}}$$

La première formule s'interprète comme suit :

– ***A $1/\sqrt{\mu}$ près la coordonnée d'un individu est égale à la moyenne arithmétique simple des coordonnées des catégories auxquelles il appartient.***

En effet $\mathbf{Xa} = \sum_{j=1}^{p}\mathbf{X}_j\mathbf{a}_j$. Pour un individu i les seuls termes non nuls sont ceux correspondant aux catégories possédées (une par variable).

La deuxième formule montre que :

– ***A $1/\sqrt{\mu}$ près la coordonnée d'une catégorie j est égale à la moyenne arithmétique des coordonnées des n_j individus de cette catégorie.***

Les points représentatifs des catégories dans les graphiques factoriels doivent donc être considérés comme des barycentres : les proximités devront être interprétées en terme de proximités entre points moyens de groupes d'individus.

On a à $1/\sqrt{\mu}$ près, la propriété des « moyennes réciproques » qui est à l'origine de certaines présentations de l'analyse des correspondances multiples (*dual scaling*).

z étant une variable de moyenne nulle il s'ensuit que pour chaque variable $\mathcal{X}_i$ les coordonnées de ses catégories (pondérées par les effectifs) sont de moyenne nulle.

Il est possible de représenter simultanément individus et catégories des variables $\mathcal{X}_i$ car les points représentatifs des catégories sont barycentres de groupes d'individus.

Nous conseillons toutefois d'utiliser le système suivant de coordonnées afin de conserver la propriété barycentrique :

$$\mathbf{z} \text{ de variance } \mu \text{ et } \boldsymbol{\alpha} = \mathbf{D}^{-1}\mathbf{X}'\mathbf{z} = \sqrt{\mu}\mathbf{a}$$

10.1.3.4 Propriétés des valeurs propres

Le rang de **X** étant $\sum_{i=1}^{p} m_i - p + 1$, si $n > \Sigma m_i$, le nombre de valeurs propres non trivialement égales à 0 ou 1 est $\sum_{i=1}^{p} m_i - p = q$.

La somme des valeurs propres non triviales vaut :

$$\boxed{\sum_{i=1}^{q} \mu_i = \frac{1}{p}\sum_{i=1}^{p} m_i - 1}$$

L'inertie est donc égale au nombre moyen de catégories diminué d'une unité : c'est une quantité qui ne dépend pas des liaisons entre les variables et n'a donc aucune signification statistique.

La moyenne des q valeurs propres vaut $1/p$. Cette quantité peut jouer dans une certaine mesure le rôle d'un seuil d'élimination pour les valeurs propres inférieures comme nous allons le voir.

La somme des carrés des valeurs propres est liée, elle, à des indices statistiques.

μ^2 étant valeur propre du carré de la matrice à diagonaliser on a :

$$1 + \sum_{i=1}^{q} (\mu_i)^2 = \text{Trace}\left(\left(\frac{1}{p}\sum_{i=1}^{p} \mathbf{A}_i\right)^2\right)$$

d'où : $$\sum(\mu_i)^2 = \frac{1}{p^2}\sum \text{Trace}\,(\mathbf{A}_i)^2 + \frac{1}{p^2}\sum_{i \neq j}\sum \text{Trace}\,(\mathbf{A}_i\mathbf{A}_j) - 1$$

comme $\mathbf{A}_i^2 = \mathbf{A}_i$: $$\sum(\mu_i)^2 = \frac{1}{p^2}\sum_i m_i + \frac{1}{p^2}\sum_{i \neq j}\sum (1 + \varphi_{ij}^2) - 1$$

$$\boxed{\sum_{i=1}^{q} (\mu_i)^2 = \frac{1}{p^2}\sum_{i=1} (m_i - 1) + \frac{1}{p^2}\sum_{i \neq j}\sum \varphi_{ij}^2}$$

où φ_{ij}^2 est le φ^2 de K. Pearson du croisement de $\mathcal{X}_i$ avec $\mathcal{X}_j$.

Si les p variables $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_p$ sont indépendantes deux à deux $\varphi_{ij}^2 = 0$ et $\sum_i (\mu_i)^2 = \frac{1}{p^2}\sum_i (m_i - 1) = \frac{1}{p^2} q$.

On a donc :
$$\frac{1}{q}\sum_i (\mu_i)^2 = \frac{1}{p^2} = \left(\frac{1}{q}\sum_i \mu_i\right)^2$$

La moyenne des carrés ne peut être égale au carré de la moyenne que si toutes les valeurs propres sont égales. Le cas de l'indépendance correspond donc à $\mu_i = \frac{1}{p}\,\forall i$.

On retrouve également cette situation si les données sont recueillies selon un plan équilibré où les $m_1 m_2 \ldots m_p$ combinaisons possibles des modalités des $\mathscr{X}_i$ sont observées avec le même effectif car tous les tableaux croisés $\mathbf{X}_i'\mathbf{X}_j$ ont alors les mêmes profils. Pour un tel plan d'expérience l'analyse des correspondances multiples est donc inutile.

10.1.3.5 AFC du tableau de Burt

Si l'on soumet le tableau $\mathbf{B}$ à une analyse des correspondances on retrouve, à une constante multiplicative près, les mêmes coordonnées factorielles des catégories.

Le tableau de Burt étant symétrique les totaux de lignes et de colonnes sont égaux (on retrouve p fois les totaux marginaux).

Le tableau des profils-lignes associées à $\mathbf{B}$ est donc $(p\mathbf{D})^{-1}\,\mathbf{B}$. Le tableau des profils-colonnes associé à $\mathbf{B}$ est $\mathbf{B}(p\mathbf{D})^{-1}$. L'AFC de $\mathbf{B}$ revient donc à diagonaliser :

$$\left(\frac{1}{p}\mathbf{D}^{-1}\mathbf{B}\right)^2$$

qui conduit aux mêmes vecteurs propres que $\frac{1}{p}\mathbf{D}^{-1}\mathbf{B}$ avec des valeurs propres égales à μ^2.

10.2 AUTRES PRÉSENTATIONS

L'extension formelle du cas $p = 2$ au cas général ne suffit pas pour conférer un statut de méthode statistique à l'analyse des correspondances multiples. Les présentations qui suivent, la reliant à d'autres méthodes, y contribuent en apportant des éclairages différents. Chacune de ces présentations correspond à une "découverte" indépendante de l'ACM.

10.2.1 Analyse canonique généralisée de p tableaux d'indicatrices

On sait que l'analyse des correspondances d'un tableau de contingence est une analyse canonique particulière, celle des tableaux $\mathbf{X}_1$ et $\mathbf{X}_2$.

Lorsqu'il y a p tableaux d'indicatrices associés à p variables qualitatives $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$, il est naturel d'utiliser la généralisation de l'analyse canonique étudiée au chapitre 8.

Celle-ci revient à chercher les vecteurs propres de la somme des opérateurs de projection sur les sous-espaces engendrés par les colonnes des $\mathbf{X}_i$.

Au coefficient $1/p$ près, les valeurs propres sont donc les mêmes qu'en analyse des correspondances multiples. Les composantes $\mathbf{z}$ sont donc identiques aux variables auxiliaires de la généralisation de Carroll de l'analyse canonique.

10.2.2 Un critère d'association maximale

Puisque l'analyse des correspondances multiples est identique à l'analyse canonique généralisée de $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$, les variables $\mathbf{z}$ rendent maximale $\sum_{i=1}^{p} R^2(\mathbf{z}\,;\mathbf{X}_i)$.

Or $\mathbf{X}_i$ étant un tableau d'indicatrices, le coefficient de corrélation multiple n'est autre que le rapport de corrélation $\eta(\mathbf{z}/\mathbf{X}_i)$ (chapitre 6, paragr. 6.4).

Les variables $\mathbf{z}$ sont donc les variables de variance μ, non corrélées deux à deux vérifiant :

$$\boxed{\max_{\mathbf{z}} \sum_{i=1}^{p} \eta^2(\mathbf{z}/\mathcal{X}_i)}$$

Si l'on se rappelle qu'en ACP normée, les composantes principales rendaient maximale $\sum_{j=1}^{p} r^2(\mathbf{c}\,;\mathbf{x}^j)$ on a ici l'équivalent d'une ACP sur variables qualitatives, la mesure de liaison étant η^2 au lieu de r^2.

L'analyse des correspondances multiples revient donc à résumer p variables qualitatives par des variables numériques de variance maximale les plus corrélées possible, au sens défini précédemment, avec les $\mathcal{X}_i$.

Lorsque les variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$ sont dichotomiques ($m_i = 2$) le tableau $\mathbf{X}$ possède $2p$ colonnes.

$$\mathbf{X} = \left[\begin{array}{c|c|c|c|c} 01 & 10 & 01 & & \\ & & & \cdots & \\ & & & & \end{array}\right]$$

Le coefficient de corrélation multiple au carré entre $\mathbf{z}$ et $\mathbf{X}_i$ est alors égal au carré du coefficient de corrélation linéaire simple entre $\mathbf{z}$ et l'une des deux indicatrices de $\mathcal{X}_i$ puisque la somme des deux indicatrices vaut 1.

On a donc : $$\sum_{i=1}^{p} \eta^2(\mathbf{z} ; \mathcal{X}_i) = \sum_{i=1}^{p} r^2(\mathbf{z} ; \mathbb{1}_{A}^{i}) = \sum_{i=1}^{p} r^2(\mathbf{z} ; \mathbb{1}_{\bar{A}}^{i})$$

Dans ce cas l'analyse des correspondances multiples de $\mathbf{X}$ revient à effectuer une ACP normée, c'est-à-dire sur la matrice de corrélation, sur un tableau à n lignes et p colonnes obtenu en ne conservant qu'une indicatrice par variable qualitative.

10.2.3 Quantification optimale de variables qualitatives

On retrouve la solution de l'analyse des correspondances multiples, tout au moins l'équation du premier facteur, en cherchant à résoudre le problème suivant : transformer de façon optimale (selon un critère à définir) chaque variable qualitative à m_i modalités en une variable discrète à m_i valeurs. On sait qu'une telle quantification s'écrit $\xi_i = \mathbf{X}_i \mathbf{a}_i$ où ξ_i est la variable numérique obtenue, $\mathbf{a}_i$ le vecteur des valeurs numériques à attribuer aux modalités.

10.2.3.1 ACP de variables quantifiées

On cherche ici à obtenir une ACP des ξ_i qui soit la meilleure possible au sens où la première valeur propre λ_1 de la matrice de corrélation des ξ_j est maximale. Ceci revient à chercher :

$$\max_{\xi_1, \xi_2, \ldots, \xi_p} \left(\max_{\mathbf{z}} \sum_{j=1}^{p} r^2(\mathbf{z} ; \xi_j) \right)$$

or : $$\max_{\xi_j} r^2(\mathbf{z} ; \xi_j) = R^2(\mathbf{z} ; \mathbf{X}_j)$$

on est donc amené à rechercher le max de $\sum_{j=1}^{p} R^2(\mathbf{z} ; \mathbf{X}_j)$. Les « codages » optimaux des catégories ne sont donc autres que les coordonnées de ces catégories sur le premier axe de l'analyse des correspondances multiples de $\mathbf{X}$.

10.2.3.2 Guttman et l'homogénéité maximale

En 1941 L. L. Guttman avait abouti aux équations de l'analyse des correspondances multiples en résolvant le problème suivant : étant donné un questionnaire à choix multiple à p questions ayant chacune m_j modalités de réponse (une seule réponse possible à chaque question), on veut attribuer des notes à chaque modalité de telle sorte que les variables numériques ainsi créées ξ_j soient les plus cohérentes au sens suivant : les réponses aux p questions doivent conduire à des notes proches, tout en donnant une note globale moyenne la plus dispersée possible.

Considérons le tableau n, p des variables ξ_j :

$$\begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ n \end{array}\begin{array}{c} \xi_1 \;\cdots\cdots\; \xi_p \\ \left[\begin{array}{ccc} \xi_{11} & \cdots\cdots & \xi_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & \cdots\cdots & \xi_{np} \end{array}\right]\end{array}$$

notons $\bar{\xi}_1, \bar{\xi}_2, \ldots, \bar{\xi}_p$ les moyennes des différentes lignes :

$$\bar{\xi}_i = \frac{1}{p}\sum_{j=1}^{p}\xi_{ij}$$

Supposons, ce qui ne nuit pas à la généralité que chaque ξ_j est une variable de moyenne nulle.

On cherche alors à avoir des mesures les plus homogènes possible en minimisant en moyenne la dispersion intra-individuelle.

Pour chaque observation celle-ci vaut $\frac{1}{p}\sum_{j=1}^{p}(\xi_{ij} - \bar{\xi}_i)^2$ donc en moyenne elle vaut :

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(\xi_{ij} - \bar{\xi}_i)^2$$

La variance totale du tableau (ξ_{ij}) étant égale à la moyenne des variances plus la variance des moyennes :

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(\xi_{ij})^2 = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(\xi_{ij} - \bar{\xi}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\bar{\xi}_i)^2$$

il revient au même de maximiser :

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(\bar{\xi}_i)^2}{\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(\xi_{ij})^2}$$

Or : $$\boldsymbol{\xi}_j = \mathbf{X}_j\mathbf{a}_j \quad \text{et} \quad \bar{\boldsymbol{\xi}} = \frac{1}{p}\sum_{j=1}^{p}\mathbf{X}_j\mathbf{a}_j = \frac{1}{p}\mathbf{Xa}$$

donc : $$\frac{1}{n}\sum_{i=1}^{n}(\bar{\xi}_i)^2 = \frac{1}{n}\left(\frac{1}{p}\mathbf{Xa}\right)'\left(\frac{1}{p}\mathbf{Xa}\right) = \frac{1}{np^2}\mathbf{a}'\mathbf{X}'\mathbf{Xa}$$

$$\sum_{i=1}^{n}(\xi_{ij})^2 = \xi_j'\xi_j = (\mathbf{X}_j\mathbf{a}_j)'(\mathbf{X}_j\mathbf{a}_j) = \mathbf{a}_j'\mathbf{D}_j\mathbf{a}_j$$

d'où :
$$\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}(\xi_{ij})^2 = \frac{1}{np}\sum_{j=1}^{p}\mathbf{a}_j'\mathbf{D}_j\mathbf{a}_j = \frac{1}{np}\mathbf{a}'\mathbf{D}\mathbf{a}$$

La quantité critère vaut donc :

$$\frac{\dfrac{1}{np^2}\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a}}{\dfrac{1}{np}\mathbf{a}'\mathbf{D}\mathbf{a}} = \frac{1}{p}\frac{\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a}}{\mathbf{a}'\mathbf{D}\mathbf{a}}$$

Son maximum est atteint pour **a** vecteur propre associé à la plus grande valeur propre λ_1 de $\frac{1}{p}\mathbf{D}^{-1}\mathbf{X}'\mathbf{X}$. On retrouve bien le premier facteur de l'ACM de **X**.

10.2.4 Approximation d'ACP non linéaire

Revenons sur le chapitre 7, § 7.6 : pour p variables numériques $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^p$, l'ACP cherche une combinaison **linéaire** de variance maximale :

$$\max V\left(\sum_{j=1}^{p} u_j \mathbf{x}^j\right)$$

Si l'on veut s'affranchir de la linéarité, on peut chercher des transformations fonctionnelles $\varphi^1(\mathbf{x}^1), \ldots, \varphi^p(\mathbf{x}^p)$ des variables telles que $V\left(\sum_{j=1}^{p}\varphi^j(\mathbf{x}^j)\right)$ soit maximal.

Choisissons pour les φ^j des fonctions en escalier (constantes par morceaux) ou splines de degré 0. On sait que ces fonctions permettent d'approximer n'importe quelle fonction numérique.

Concrètement on découpera l'intervalle de variation de $\mathbf{x}^j$ en m_j classes (fig. 10.1).

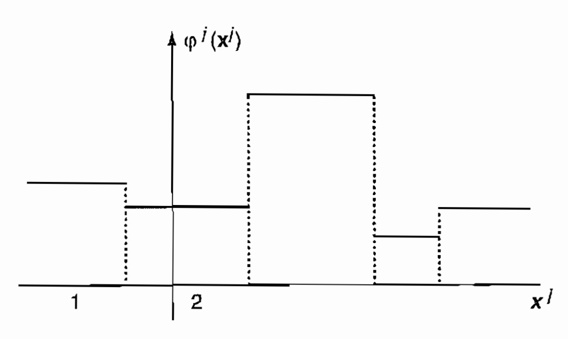


FIGURE 10.1

$\varphi^j(\mathbf{x}^j)$ est une fonction prenant les valeurs $a_1, a_2, \ldots, a_{mj}$ sur les intervalles de découpage ; elle s'explicite comme la combinaison linéaire des variables indicatrices des classes du découpage, à coefficients $a_1, a_2, \ldots, a_{m_j}$.

Le critère max $V\left(\sum_j \varphi^j(\mathbf{x}^j)\right)$ est donc identique au critère max $V\left(\sum_j \mathbf{X}_j\mathbf{a}_j\right)$. La solution est alors donnée par la première composante de l'analyse des correspondances multiples du tableau **X** obtenu en découpant en classes chacune des variables numériques.

La pratique qui consiste à découper en classes des variables numériques, donc à les rendre qualitatives, pour ensuite effectuer une analyse des correspondances multiples se justifie par le fait qu'il s'agit d'une analyse non linéaire des données.

Sous réserve d'avoir suffisamment d'observations par classe on peut ainsi visualiser des liaisons non linéaires entre variables qui ne seraient pas apparues en ACP ordinaire où l'on travaille avec la matrice **R** des corrélations linéaires.

10.3 PRATIQUE DE L'ANALYSE DES CORRESPONDANCES MULTIPLES

L'interprétation des résultats d'une ACM se fait *grosso-modo* comme en analyse des correspondances sur tableau de contingence et comme en ACP. On prendra garde ici au fait que les pourcentages d'inertie n'ont qu'un intérêt restreint. La sélection et l'interprétation des axes factoriels se fera essentiellement à l'aide des contributions des variables actives et des valeur-tests associées aux variables supplémentaires. Rappelons une fois encore la signification des proximités entre points-colonnes sur un plan factoriel : il s'agit d'une proximité, en projection, de points moyens de catégories représentant plusieurs individus.

10.3.1 Les contributions

10.3.1.1 Contributions à un axe factoriel

Une catégorie d'effectif n_j qui a une coordonnée a_j sur un axe factoriel fournit une contribution (CTR) égale à :

$$\text{CTR}(j) = \frac{\dfrac{n_j}{np}(a_j)^2}{\mu}$$

On repèrera les modalités intéressantes qui ont une contribution supérieure à leur poids $\dfrac{n_j}{np}$.

En correspondances multiples, les modalités d'une même variable $\mathcal{X}_i$ ont des contributions qui peuvent être cumulées.

On définit la contribution cumulée de $\mathscr{X}_i$ comme :

$$\mathrm{CTR}(\mathscr{X}_i) = \sum_{j=1}^{m_i} \mathrm{CTR}(j) = \frac{1}{\mu} \sum_{j=1}^{m_i} \frac{n_j}{np}(a_j)^2$$

a_j étant à $\sqrt{\mu}$ près la moyenne des coordonnées des individus de la catégorie j de $\mathscr{X}_{ui}$, les contributions cumulées sont reliées au rapport de corrélation entre la composante $\mathbf{z}$ de variance μ et la variable $\mathscr{X}_i$ par :

$$\eta^2(\mathbf{z}/\mathscr{X}_i) = p\mu \, \mathrm{CTR}(\mathscr{X}_i)$$

Remarquons que $\eta^2 < 1$ entraîne $\mathrm{CTR}(\mathscr{X}_i) < \dfrac{1}{p\mu}$ et que $\dfrac{1}{p} \sum_{i=1}^{p} \eta^2(\mathbf{z}/\mathscr{X}_i) = \mu$.

On utilise comme en ACP les contributions des individus $\dfrac{1}{n}(z_i)^2/\mu$, et comme en ACP et en AFC les cosinus carrés avec les axes pour juger de la qualité d'une projection.

On pourra utiliser ici le résultat donné au chapitre 7 : un individu aura une contribution significative si celle-ci dépasse 3.84 fois son poids.

10.3.1.2 Contributions à l'inertie totale

L'inertie totale vaut, rappelons-le, $\dfrac{1}{p} \sum_{i=1}^{p} m_i - 1$. Le nuage des profils-colonnes a pour centre de gravité le vecteur de $\mathbb{R}^n$ dont toutes les composantes valent $1/n$: en effet la somme des colonnes du tableau disjonctif est le vecteur constant dont toutes les composantes valent p.

La métrique du χ^2 pour le nuage des profils-colonnes est donc la métrique diagonale $n\mathbf{I}_n$ (diagonale des inverses des fréquences marginales).

Le carré de distance d'un point catégorie $\mathbf{j}$ au centre de gravité $\mathbf{g}$ vaut donc :

$$d^2(\mathbf{j} ; \mathbf{g}) = n \sum_{i=1}^{n} (x_{ij}/n_j - 1/n)^2$$

où x_{ij} est le terme courant de la $j^{\text{ème}}$ colonne du tableau disjonctif. Comme $x_{ij} = 0$ ou 1 on a $x_{ij}^2 = x_{ij}$ d'où :

$$d^2(\mathbf{j}, \mathbf{g}) = n \sum_{i=1}^{n} \left(\frac{x_{ij}}{n_j^2} + \frac{1}{n^2} - \frac{2x_{ij}}{nn_j} \right)$$

comme $\sum_i x_{ij} = n_j$ il vient : $$d^2(\mathbf{j}, \mathbf{g}) = \frac{n}{n_j} - 1$$

Une catégorie est donc d'autant plus éloignée du centre que son effectif est faible.

Son inertie vaut $\frac{n_j}{np} d^2(\mathbf{j}, \mathbf{g}) = \left(1 - \frac{n_j}{n}\right)\frac{1}{p}$:

$$I(j) = \frac{1}{p}\left(1 - \frac{n_j}{n}\right)$$

La contribution d'une modalité à l'inertie est fonction décroissante de son effectif. Il convient donc d'éviter de travailler avec des catégories d'effectif trop faible, qui risquent de perturber les résultats de l'analyse (absence de robustesse).

L'inertie totale d'une variable, $I(\mathcal{X}_i)$, vaut :

$$\sum_{j=1}^{m_i}\left(1 - \frac{n_j}{n}\right)\frac{1}{p} = \frac{(m_i - 1)}{p}$$

$$I(\mathcal{X}_i) = \frac{(m_i - 1)}{p}$$

sa contribution est donc : $$\frac{I(\mathcal{X}_i)}{\frac{1}{p}\sum_i m_i - 1} = \frac{m_i - 1}{\sum_i (m_i - 1)}$$

Elle est d'autant plus importante que son nombre de catégories est élevé. On recommande généralement pour cette raison d'éviter des disparités trop grandes entre les nombres de catégories des variables $\mathcal{X}_i$, lorsque l'on a le choix du découpage.

10.3.2 L'usage de variables supplémentaires

Déjà évoqué lors de l'étude de l'ACP, l'usage de variables supplémentaires est très courant en analyse des correspondances multiples.

Rappelons que les variables actives sont celles qui déterminent les axes. Les variables supplémentaires ne participent pas au calcul des valeurs propres et vecteurs propres mais peuvent être représentées sur les plans factoriels selon le principe barycentrique pour les variables qualitatives : chaque catégorie est le point-moyen d'un groupe d'individus.

Pour les catégories des variables suplémentaires qualitatives on calcule comme en ACP des valeurs-test mesurant en nombre d'écart-type l'éloignement de l'origine.

Enfin il est possible de mettre en variables supplémentaires les variables numériques qui ne peuvent pas être actives (à moins de les rendre qualitatives par découpage en classes) : Elles peuvent être positionnées dans un cercle de corrélation avec pour coordonnées les corrélations avec les composantes de l'analyse.

Soit a_j la coordonnée d'une catégorie d'une variable supplémentaire, d'effectif n_j, sur un certain axe d'inertie égale à μ :

On sait que si les n_j individus de cette catégorie étaient pris au hasard parmi les n individus de l'échantillon (sans remise) la moyenne des coordonnées des n_j individus concernés serait une variable aléatoire centrée (puisque par construction les composantes **z** sont de moyenne nulle) et de variance égale à $\frac{\mu}{n_j}\frac{n-n_j}{n-1}$ (voir chapitre 7).

Avec les conventions habituelles de la représentation simultanée a_j est égale à $1/\sqrt{\mu}$ fois la moyenne des coordonnées, la quantité $a_j\sqrt{n_j}\sqrt{\frac{n-1}{n-n_j}}$ est donc la valeur-test.

Le calcul des valeurs-test n'est légitime que pour des variables supplémentaires n'ayant pas servi à la détermination des axes. Leur utilisation pour des variables actives ne doit être considérée qu'à titre indicatif : les valeurs-test pour les variables actives sout en général très élevées, ce qui est normal car les variables actives déterminent les axes.

10.4 UN EXEMPLE : LES RACES CANINES

Les données communiquées par M. Tenenhaus (tableau 10.1) décrivent les caractéristiques de 27 races de chiens au moyen de variables qualitatives, les 6 premières ont été considérées comme actives, la septième, « fonction », comme supplémentaire : ses trois modalités sont « compagnie » « chasse » « utilité ».

On remarquera que les paires d'individus (5, 26) (8, 22) (11, 19) ont des valeurs identiques pour les 7 variables, il y aura donc des observations confondues.

Le nombre de modalités actives est 16, ce qui conduit à 10 facteurs et à une inertie totale de $\frac{16}{6}-1=\frac{5}{3}\simeq 1.667$, le critère $\mu > 1/p$ conduit à ne retenir que trois axes, le diagramme des valeurs propres montre cependant une chute après μ_2. On interprètera donc uniquement les deux premiers axes (tableau 10.2)[1].

L'axe 1 oppose (à droite) les chiens de petite taille, affectueux, qui coïncident avec les chiens de compagnie (valeur-test 4.06), aux chiens de grande taille, très rapides et agressifs (fonction « utilité »). L'axe 2 oppose (en bas) les chiens de chasse, de taille moyenne, très intelligents à des chiens lents et peu intelligents.

Le tableau 10.3 est le tableau de Burt qui résume les liaisons deux à deux entre les 6 variables actives.

Le tableau 10.4 permet de repérer les modalités contribuant fortement à l'inertie des axes et sa lecture doit être complétée par celle du tableau 10.5 qui fournit les valeurs tests.

Le tableau 10.6 permet d'apprécier la qualité de la représentation graphique (fig. 10.2).

1 Les calculs ont été effectués à l'aide du logiciel SPAD, version 5.6.

Tableau 10.1
Tableau disjonctif

		1			2			3			4			5		6		7		
		Taille			Poids			Vélocité			Intelligence			Affection		Agressivité		Fonction		
		−	+	++	−	+	++	−	+	++	−	+	++	−	+	−	+	Co.	Ch.	Ut.
1	Beauceron	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0	1
2	Basset	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0
3	Berger Allemand	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0	1
4	Boxer	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	1	0	0
5	Bull-Dog	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	1	0	0
6	Bull-Mastiff	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	1
7	Caniche	1	0	0	1	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0
8	Chihuahua	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0	1	0	0
9	Cocker	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0
10	Colley	0	0	1	0	1	0	0	0	1	0	1	0	0	1	1	0	1	0	0
11	Dalmatien	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	0
12	Dobermann	0	0	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0	0	1
13	Dogue Allemand	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1
14	Épagneul Breton	0	1	0	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1	0
15	Épagneul Français	0	0	1	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0
16	Fox-Hound	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	0	1	0
17	Fox-Terrier	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	1	1	0	0
18	Grand Bleu de Gascogne	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	1	0	1	0
19	Labrador	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0
20	Lévrier	0	0	1	0	1	0	0	0	1	1	0	0	1	0	1	0	0	1	0
21	Mastiff	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0	1
22	Pékinois	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0	1	0	0
23	Pointer	0	0	1	0	1	0	0	0	1	0	0	1	1	0	1	0	0	1	0
24	Saint-Bernard	0	0	1	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0	1
25	Setter	0	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0
26	Teckel	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	1	0	0
27	Terre-Neuve	0	0	1	0	0	1	1	0	0	0	1	0	1	0	1	0	0	0	1

TABLEAU 10.2

DIAGRAMME DES 10 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	0.4816	28.90	28.90	**
2	0.3847	23.08	51.98	**
3	0.2110	12.66	64.64	************************************
4	0.1576	9.45	74.09	***************************
5	0.1501	9.01	83.10	*************************
6	0.1233	7.40	90.50	*********************
7	0.0815	4.89	95.38	**************
8	0.0457	2.74	98.12	********
9	0.0235	1.41	99.54	****
10	0.0077	0.46	100.00	**

Tableau 10.3

TABLEAU DE BURT

	TA1	TA2	TA3	PO1	PO2	PO3	VE1	VE2	VE3	IN1	IN2	IN3	AF1	AF2	AG1	AG2
TA1	7	0	0													
TA2	0	5	0													
TA3	0	0	15													
PO1	7	1	0	8	0	0										
PO2	0	4	10	0	14	0										
PO3	0	0	5	0	0	5										
VE1	5	1	4	6	0	4	10	0	0							
VE2	2	4	2	2	6	0	0	8	0							
VE3	0	0	9	0	8	1	0	0	9							
IN1	3	0	5	3	3	2	4	1	3	8	0	0				
IN2	3	4	6	4	7	2	5	5	3	0	13	0				
IN3	1	1	4	1	4	1	1	2	3	0	0	6				
AF1	1	0	12	1	7	5	5	2	6	6	4	3	13	0		
AF2	6	5	3	7	7	0	5	6	3	2	9	3	0	14		
AG1	5	3	6	5	8	1	5	5	4	3	8	3	5	9	14	0
AG2	2	2	9	3	6	4	5	3	5	5	5	3	8	5	0	13
	TA1	TA2	TA3	PO1	PO2	PO3	VE1	VE2	VE3	IN1	IN2	IN3	AF1	AF2	AG1	AG2

TABLEAU 10.4

MODALITES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN-LIBELLE	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1 . Taille																	
TA1 - PETITE TAILLE	4.32	2.86	-1.18	0.92	-0.62	0.12	-0.02	12.6	9.6	7.8	0.4	0.0	0.49	0.30	0.13	0.01	0.00
TA2 - TAILLE MOYENNE	3.09	4.40	-0.85	-1.23	1.02	0.34	-0.31	4.6	12.2	15.1	2.3	2.0	0.16	0.34	0.23	0.03	0.02
TA3 - GRANDE TAILLE	9.26	0.80	0.84	-0.02	-0.05	-0.17	0.11	13.5	0.0	0.1	1.7	0.8	0.88	0.00	0.00	0.04	0.02
			CONTRIBUTION CUMULEE =					30.7	21.8	23.0	4.4	2.8					
2 . Poids																	
PO1 - PETIT POIDS	4.94	2.38	-1.17	0.82	-0.36	0.16	-0.05	14.0	8.7	3.0	0.9	0.1	0.58	0.29	0.05	0.01	0.00
PO2 - POIDS MOYEN	8.64	0.93	0.31	-0.82	-0.23	-0.12	-0.19	1.7	15.1	2.2	0.8	2.1	0.10	0.72	0.06	0.02	0.04
PO3 - POIDS ELEVE	3.09	4.40	1.02	0.97	1.22	0.07	0.61	6.6	7.6	21.8	0.1	7.8	0.23	0.22	0.34	0.00	0.09
			CONTRIBUTION CUMULEE =					22.3	31.4	27.0	1.7	9.9					
3 . Vélocité																	
VE1 - LENT	6.17	1.70	-0.32	1.04	0.40	-0.08	0.31	1.3	17.5	4.7	0.3	3.8	0.06	0.64	0.09	0.00	0.06
VE2 - ASSEZ RAPIDE	4.94	2.38	-0.60	-0.89	0.36	0.37	-0.37	3.7	10.1	3.0	4.3	4.5	0.15	0.33	0.05	0.06	0.06
VE3 - TRES RAPIDE	5.56	2.00	0.89	-0.37	-0.76	-0.24	-0.01	9.2	2.0	15.3	2.0	0.0	0.40	0.07	0.29	0.03	0.00
			CONTRIBUTION CUMULEE =					14.2	29.6	23.0	6.6	8.4					
4 . Intelligence																	
IN1 - PEU INTELLIGENT	4.94	2.38	0.35	0.81	-0.35	0.02	-1.04	1.2	8.4	2.9	0.0	35.2	0.05	0.28	0.05	0.00	0.45
IN2 - INTELLIGENCE MOYENNE	8.02	1.08	-0.37	-0.29	0.49	-0.60	0.15	2.3	1.7	9.3	18.5	1.1	0.13	0.08	0.23	0.34	0.02
IN3 - TRES INTELLIGENT	3.70	3.50	0.34	-0.46	-0.60	1.28	1.06	0.9	2.0	6.3	38.2	27.9	0.03	0.06	0.10	0.46	0.32
			CONTRIBUTION CUMULEE =					4.4	12.1	18.5	56.8	64.3					
5 . Affection																	
AF1 - PEU AFFECTUEUX	8.02	1.08	0.84	0.29	0.07	-0.08	-0.04	11.6	1.7	0.2	0.4	0.1	0.65	0.08	0.00	0.01	0.00
AF2 - AFFECTUEUX	8.64	0.93	-0.78	-0.27	-0.06	0.08	0.04	10.8	1.6	0.2	0.3	0.1	0.65	0.08	0.00	0.01	0.00
			CONTRIBUTION CUMULEE =					22.4	3.3	0.3	0.7	0.2					
6 . Agressivité																	
AG1 - PEU AGRESSIF	8.64	0.93	-0.40	-0.19	-0.31	-0.51	0.35	2.9	0.8	3.9	14.4	7.0	0.17	0.04	0.10	0.28	0.13
AG2 - AGRESSIF	8.02	1.08	0.43	0.21	0.33	0.55	-0.37	3.1	0.9	4.2	15.5	7.5	0.17	0.04	0.10	0.28	0.13
			CONTRIBUTION CUMULEE =					6.0	1.8	8.2	29.8	14.5					

TABLEAU 10.5

MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
1 . Taille													
TA1 - PETITE TAILLE	7	7.00	-3.6	2.8	-1.9	0.4	-0.1	-1.18	0.92	-0.62	0.12	-0.02	2.86
TA2 - TAILLE MOYENNE	5	5.00	-2.1	-3.0	2.5	0.8	-0.8	-0.85	-1.23	1.02	0.34	-0.31	4.40
TA3 - GRANDE TAILLE	15	15.00	4.8	-0.1	-0.3	-1.0	0.6	0.84	-0.02	-0.05	-0.17	0.11	0.80
2 . Poids													
PO1 - PETIT POIDS	8	8.00	-3.9	2.7	-1.2	0.5	-0.2	-1.17	0.82	-0.36	0.16	-0.05	2.38
PO2 - POIDS MOYEN	14	14.00	1.6	-4.3	-1.2	-0.6	-1.0	0.31	-0.82	-0.23	-0.12	-0.19	0.93
PO3 - POIDS ELEVE	5	5.00	2.5	2.4	3.0	0.2	1.5	1.02	0.97	1.22	0.07	0.61	4.40
3 . Vélocité													
VE1 - LENT	10	10.00	-1.3	4.1	1.6	-0.3	1.2	-0.32	1.04	0.40	-0.08	0.31	1.70
VE2 - ASSEZ RAPIDE	8	8.00	-2.0	-2.9	1.2	1.2	-1.2	-0.60	-0.89	0.36	0.37	-0.37	2.38
VE3 - TRES RAPIDE	9	9.00	3.2	-1.3	-2.8	-0.9	0.0	0.89	-0.37	-0.76	-0.24	-0.01	2.00
4 . Intelligence													
IN1 - PEU INTELLIGENT	8	8.00	1.2	2.7	-1.2	0.1	-3.4	0.35	0.81	-0.35	0.02	-1.04	2.38
IN2 - INTELLIGENCE MOYENNE	13	13.00	-1.8	-1.4	2.4	-3.0	0.7	-0.37	-0.29	0.49	-0.60	0.15	1.08
IN3 - TRES INTELLIGENT	6	6.00	0.9	-1.3	-1.6	3.5	2.9	0.34	-0.46	-0.60	1.28	1.06	3.50
5 . Affection													
AF1 - PEU AFFECTUEUX	13	13.00	4.1	1.4	0.3	-0.4	-0.2	0.84	0.29	0.07	-0.08	-0.04	1.08
AF2 - AFFECTUEUX	14	14.00	-4.1	-1.4	-0.3	0.4	0.2	-0.78	-0.27	-0.06	0.08	0.04	0.93
6 . Agressivité													
AG1 - PEU AGRESSIF	14	14.00	-2.1	-1.0	-1.6	-2.7	1.8	-0.40	-0.19	-0.31	-0.51	0.35	0.93
AG2 - AGRESSIF	13	13.00	2.1	1.0	1.6	2.7	-1.8	0.43	0.21	0.33	0.55	-0.37	1.08
7 . Fonction													
FO1 - COMPAGNIE	10	10.00	-4.1	0.4	-0.3	-0.3	0.0	-1.04	0.10	-0.07	-0.09	-0.01	1.70
FO2 - CHASSE	9	9.00	1.2	-1.6	-1.3	-0.7	-1.6	0.32	-0.43	-0.35	-0.18	-0.44	2.00
FO3 - UTILITAIRE	8	8.00	3.1	1.2	1.6	1.1	1.7	0.94	0.37	0.48	0.32	0.51	2.38

Tableau 10.6

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
BEAUCERON	3.70	1.14	0.32	-0.42	-0.10	-0.21	-0.12	0.8	1.7	0.2	1.1	0.3	0.09	0.15	0.01	0.04	0.01
BASSET	3.70	1.91	-0.25	1.10	-0.19	0.29	-0.52	0.5	11.7	0.6	2.0	6.8	0.03	0.63	0.02	0.04	0.14
BERGER ALLEMAND	3.70	1.54	0.49	-0.46	-0.50	0.58	0.28	1.8	2.1	4.4	7.8	1.9	0.15	0.14	0.16	0.22	0.05
BOXER	3.70	1.80	-0.45	-0.88	0.69	0.26	-0.46	1.5	7.5	8.4	1.6	5.1	0.11	0.43	0.27	0.04	0.12
BULL-DOG	3.70	1.64	-1.01	0.55	-0.16	-0.35	0.33	7.9	2.9	0.5	2.9	2.7	0.62	0.18	0.02	0.07	0.07
BULL-MASTIFF	3.70	2.09	0.75	0.55	0.50	0.66	0.72	4.4	2.9	4.3	10.1	12.9	0.27	0.14	0.12	0.21	0.25
CANICHE	3.70	2.16	-0.91	-0.02	-0.58	0.63	0.43	6.4	0.0	5.8	9.3	4.6	0.39	0.00	0.15	0.18	0.09
CHIHUAHUA	3.70	1.86	-0.84	0.84	-0.47	-0.09	-0.18	5.4	6.9	3.9	0.2	0.8	0.38	0.38	0.12	0.00	0.02
COCKER	3.70	1.93	-0.73	0.08	0.66	0.19	-0.10	4.1	0.1	7.7	0.8	0.3	0.28	0.00	0.23	0.02	0.01
COLLEY	3.70	1.11	0.12	-0.53	-0.33	-0.66	0.19	0.1	2.7	2.0	10.2	0.9	0.01	0.25	0.10	0.39	0.03
DALMATIEN	3.70	1.77	-0.65	-0.99	0.46	-0.19	-0.14	3.2	9.4	3.7	0.8	0.5	0.24	0.55	0.12	0.02	0.01
DOBERMANN	3.70	1.56	0.87	-0.32	-0.45	0.51	0.24	5.9	1.0	3.6	6.1	1.4	0.49	0.06	0.13	0.17	0.04
DOGUE ALLEMAND	3.70	1.95	1.05	0.51	0.17	0.06	-0.32	8.4	2.5	0.5	0.1	2.5	0.56	0.13	0.01	0.00	0.05
EPAGNEUL BRETON	3.70	2.18	-0.48	-1.04	0.06	0.60	0.25	1.8	10.4	0.1	8.5	1.5	0.10	0.49	0.00	0.17	0.03
EPAGNEUL FRANCAIS	3.70	1.20	0.14	-0.52	0.12	-0.47	0.00	0.2	2.6	0.2	5.2	0.0	0.02	0.22	0.01	0.18	0.00
FOX-HOUND	3.70	1.38	0.88	0.03	-0.36	-0.02	-0.66	5.9	0.0	2.3	0.0	10.8	0.56	0.00	0.10	0.00	0.32
FOX-TERRIER	3.70	1.78	-0.88	0.14	0.05	0.29	-0.27	6.0	0.2	0.1	1.9	1.8	0.44	0.01	0.00	0.05	0.04
GRAND BLEU DE GASCOGNE	3.70	1.44	0.52	-0.11	0.04	0.24	-0.82	2.1	0.1	0.0	1.4	16.5	0.19	0.01	0.00	0.04	0.46
LABRADOR	3.70	1.77	-0.65	-0.99	0.46	-0.19	-0.14	3.2	9.4	3.7	0.8	0.5	0.24	0.55	0.12	0.02	0.01
LEVRIER	3.70	1.35	0.68	-0.08	-0.60	-0.46	-0.35	3.5	0.1	6.2	5.0	3.1	0.34	0.01	0.26	0.16	0.09
MASTIFF	3.70	1.90	0.76	0.89	0.59	0.13	-0.18	4.4	7.6	6.1	0.4	0.8	0.30	0.41	0.18	0.01	0.02
PEKINOIS	3.70	1.86	-0.84	0.84	-0.47	-0.09	-0.18	5.4	6.9	3.9	0.2	0.8	0.38	0.38	0.12	0.00	0.02
POINTER	3.70	1.54	0.67	-0.42	-0.69	0.06	0.55	3.5	1.7	8.3	0.1	7.5	0.29	0.12	0.31	0.00	0.20
SAINT-BERNARD	3.70	1.69	0.58	0.59	0.89	-0.13	0.33	2.6	3.4	14.0	0.4	2.6	0.20	0.21	0.47	0.01	0.06
SETTER	3.70	1.14	0.50	-0.38	-0.29	-0.73	0.16	2.0	1.4	1.5	12.4	0.6	0.22	0.13	0.07	0.46	0.02
TECKEL	3.70	1.64	-1.01	0.55	-0.16	-0.35	0.33	7.9	2.9	0.5	2.9	2.7	0.62	0.18	0.02	0.07	0.07
TERRE-NEUVE	3.70	1.66	0.38	0.49	0.66	-0.58	0.64	1.1	2.3	7.7	7.9	10.0	0.09	0.14	0.26	0.20	0.24

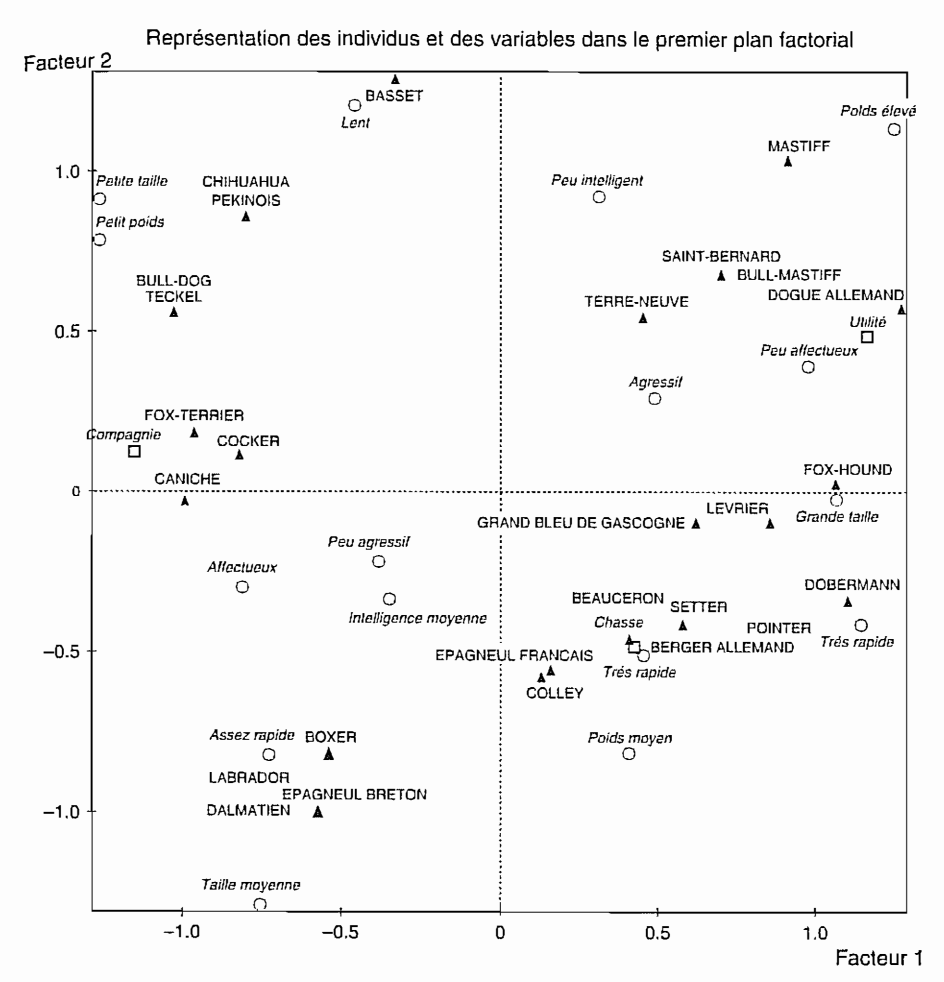

Représentation des individus et des variables dans le premier plan factorial
Facteur 2
Facteur 1
1.0
0.5
0
−0.5
−1.0
−1.0
−0.5
0
0.5
1.0
BASSET
Lent
Polds élevé
MASTIFF
Petite taille
CHIHUAHUA
PEKINOIS
Peu intelligent
Petit poids
SAINT-BERNARD
BULL-MASTIFF
BULL-DOG
TECKEL
TERRE-NEUVE
DOGUE ALLEMAND
Utilité
Peu affectueux
Agressif
FOX-TERRIER
Compagnie
COCKER
CANICHE
FOX-HOUND
LEVRIER
Grande taille
GRAND BLEU DE GASCOGNE
Peu agressif
Affectueux
DOBERMANN
BEAUCERON
SETTER
Intelligence moyenne
Chasse
POINTER
Très rapide
BERGER ALLEMAND
EPAGNEUL FRANCAIS
Très rapide
COLLEY
Assez rapide
BOXER
Poids moyen
LABRADOR
EPAGNEUL BRETON
DALMATIEN
Taille moyenne

Figure 10.2

11 Méthodes de classification

Le but des méthodes de classification est de construire une partition, ou une suite de partitions emboîtées, d'un ensemble d'objets dont on connaît les distances deux à deux. Les classes formées doivent être le plus homogène possible.*

11.1 GÉNÉRALITÉS

11.1.1 Distances et dissimilarités

En classification, que les données se présentent initialement sous forme d'un tableau individus-variables ou non, toute l'information utile est contenue dans un tableau $n \times n$ donnant les dissemblances entre les n individus à classer.

11.1.1.1 Définitions

Notons E l'ensemble des n objets à classer. Une distance est une application de $E \times E$ dans $\mathbb{R}^+$ telle que :

$$\begin{cases} d(i,j) = d(j,i) \\ d(i,j) \geq 0 \\ d(i,j) = 0 \Leftrightarrow i = j \\ d(i,j) \leq d(i,k) + d(k,j) \end{cases}$$

Rappelons que toute distance n'est pas euclidienne ; il faut pour cela qu'elle soit engendrée par un produit scalaire (voir chapitre 7).

Lorsque l'on a seulement :

$$\begin{cases} d(i,j) = d(j,i) \\ d(i,j) \geq 0 \qquad d(i,i) = 0 \end{cases}$$

on parle de dissimilarité. Une similarité est une application s telle que :

$$\begin{cases} s(i,j) = s(j,i) \\ s(i,j) \geq 0 \\ s(i,i) \geq s(i,j) \end{cases}$$

* Le lecteur désireux d'approfondir ce chapitre se reportera avec profit au livre de Nakache et Confais, 2004.

Lorsque l'on a seulement une information du type suivant : i et j sont plus semblables que k et l, on parle de préordonnance ; il s'agit d'un préordre total sur les paires d'éléments de E.

Shepard, puis Benzécri, ont montré que la connaissance d'une préordonnance suffit à reconstituer une figure géométrique de n points dans un espace euclidien de dimension réduite. Benzécri a donné la formule approchée suivante pour reconstituer une distance d_{ij} connaissant seulement le rang de cette distance parmi les $n(n-1)/2$ possibles :

$$P\left(\chi_p^2 < \frac{1}{2} d_{ij}^2\right) = \frac{2r_{ij} - 1}{n(n-1)}$$

où p est la dimension de l'espace.

Lorsque les données se présentent sous forme d'un tableau $\mathbf{X}$ de p caractères numériques, on utilise souvent la métrique euclidienne classique $\mathbf{M} = \mathbf{I}$, ou $\mathbf{M} = \mathbf{D}_{1/s^2}$, la métrique de Mahalanobis $\mathbf{M} = \mathbf{V}^{-1}$, la distance $L_1 : d(i,j) = \sum_k |x_i^k - x_j^k|$, la distance de Minkowski $L_q : d(i,j) = \left(\sum_k (x_i^k - x_j^k)^q\right)^{1/q}$.

11.1.1.2 Similarités entre objets décrits par des variables binaires

Ce cas très fréquent concerne des données du type suivant : n individus sont décrits par la présence ou l'absence de p caractéristiques. De nombreux indices de similarité ont été proposés qui combinent de diverses manières les quatre nombres suivants associés à un couple d'individus :

a = nombre de caractéristiques communes ;
b = nombre de caractéristiques possédées par i et pas par j ;
c = nombre de caractéristiques possédées par j et pas par i ;
d = nombre de caractéristiques que ne possèdent ni i, ni j.

Bien que posséder une caractéristique ou ne pas posséder la caractéristique contraire soient logiquement équivalent, a et d ne jouent pas le même rôle pour des données réelles : le fait que deux végétaux ne poussent pas dans la même région ne les rend pas nécessairement semblables.

Les indices suivants compris entre 0 et 1 sont aisément transformables en dissimilarité par complémentation à 1 :

Jaccard : $\dfrac{a}{a+b+c}$;

Dice ou Czekanowski : $\dfrac{2a}{2a+b+c}$;

Ochiaï : $\dfrac{a}{\sqrt{(a+b)(a+c)}}$;

Russel et Rao : $\dfrac{a}{a+b+c+d}$;

Rogers et Tanimoto : $\dfrac{a+d}{a+d+2(b+c)}$.

De nombreux autres indices ont été proposés.

11.1.1.3 Accord entre distances et dissimilarités

Deux distances ou dissimilarités s'accordent d'autant mieux qu'elles respectent les ordres entre proximités. A toute distance d correspond un ordre sur les parties d'éléments de E définies par des relations du type $d(a, b) \leq d(c, d)$. Pour comparer deux distances d_1 et d_2, on formera tous les quadruplets possibles de points de E et on comptera le nombre d'inégalités modifiées (ceci constitue une distance entre classes de fonctions de $E \times E$ dans $\mathbb{R}^+$ définies à un automorphisme croissant près).

L'ordre sur les paires défini par une distance s'appelle une ordonnance. Si J désigne l'ensemble des paires de E, cette ordonnance peut être représentée par un graphe sur J, c'est-à-dire une partie E de $J \times J$. Le nombre des inégalités modifiées n'est autre que le cardinal de la différence symétrique des graphes G_1 et G_2 associés à d_1 et d_2 $d(d_1 ; d_2) =$ card $(G_1 \Delta G_2)$.

11.1.2 Accord entre partitions, indice de Rand

Une partition définit une variable qualitative dont les catégories sont les classes de la partition. On pourrait donc comparer deux partitions P_1 et P_2 en étudiant le croisement des deux variables qualitatives associées. Cependant, la numérotation des classes étant arbitraire, il est préférable de considérer les paires d'individus afin de savoir si quand deux individus font partie de la même classe de P_1, ils sont dans une même classe de P_2.

11.1.2.1 Tableau des comparaisons par paires associé à une partition

On notera $\mathbf{C}$ le tableau de taille n, telle que $c_{ij} = 1$ si les individus i et j font partie de la même classe, $c_{ij} = 0$ sinon. Il est facile de voir que $\mathbf{C} = \mathbf{XX}'$ où $\mathbf{X}$ est le tableau disjonctif associé à une partition P Les c_{ij} vérifient des relations particulières puisqu'une partition est une relation d'équivalence :

$$\begin{cases} \text{Réflexivité : } c_{ii} = 1 \\ \text{Symétrie : } c_{ij} = c_{ji} \\ \text{Transitivité : } c_{ij} + c_{jk} - c_{ik} \leq 1 \end{cases}$$

La dernière relation peut ne pas sembler naturelle, mais elle traduit linéairement le fait que, si i et j sont dans une même classe, j et k dans une même classe, alors les 3 éléments sont dans la même classe.

On a de plus les formules suivantes :

- Le nombre m de classes de la partition est tel que :

$$m = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} c_{ij}}$$

- Si $n_{u.}$ désigne le nombre d'individus de la classe u :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} = Trace(\mathbf{CC}') = \sum_{u=1}^{m} n_{u.}^2$$

11.1.2.2 Accord entre deux partitions

Considérons les n^2 paires d'individus, on notera :

a : le nombre de paires dans une même classe de P_1 et dans une même classe de P_2 (accords positifs)

b : le nombre de paires dans une même classe de P_1 et séparées dans P_2

c : le nombre de paires séparées dans P_1 et dans une même classe de P_2

d : le nombre de paires séparées dans P_1 et séparées dans P_2 (accords négatifs)

Le pourcentage de paires concordantes a/n^2 est un coefficient semblable à celui de Kendall pour les ordres, mais il est plus courant d'utiliser $(a + d)/n^2$ si l'on donne la même importance à l'appartenance au complémentaire d'une classe.

On a ainsi défini le coefficient de Rand R dans la version donnée par Marcotorchino et Michaud (n^2 paires au lieu de $n(n - 1)/2$ paires dans la version originale de Rand).

En notant $\mathbf{C}^1$ et $\mathbf{C}^2$ les deux matrices de comparaisons par paire, on trouve facilement :

$$a = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^1 c_{ij}^2 = Trace(\mathbf{C}^1\mathbf{C}^2) = \sum_{u=1}^{m_1} \sum_{v=1}^{m_2} n_{uv}$$

où n_{uv} est le terme général du tableau de contingence $\mathbf{X}_1'\mathbf{X}_2$ croisant les deux partitions.

On a :

$$d = \sum_{i=1}^{n} \sum_{j=1}^{n} (1 - c_{ij}^1)(1 - c_{ij}^2)$$

Le coefficient de Rand vaut alors :

$$R = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^1 c_{ij}^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} (1 - c_{ij}^1)(1 - c_{ij}^2)}{n^2} = \frac{2\sum_u\sum_v n_{uv}^2 - \sum_u n_{u.}^2 - \sum_v n_{.v}^2 + n^2}{n^2}$$

Il prend ses valeurs entre 0 et 1 ; il est égal à 1 lorsque les deux partitions sont identiques.

La version suivante (correction de Hubert et Arabie) est également utilisée :

$$RC = \frac{n^2 \sum_{u,v} n_{uv}^2 - \sum_u n_{u.}^2 \sum_v n_{.v}^2}{\frac{1}{2} n^2 \left(\sum_u n_{u.}^2 + \sum_v n_{.v}^2 \right) - \sum_u n_{u.}^2 \sum_v n_{.v}^2}$$

Son avantage est que son espérance est nulle si les deux partitions sont indépendantes, mais l'inconvénient est que l'on peut avoir des valeurs négatives.

Exemple : Considérons les deux partitions $P_1 = (ab)(cd)$ et $P_2 = (a)(bcd)$ de 4 objets. On a :

$$\mathbf{C}^1 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad \mathbf{C}^2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

Le tableau croisé est :

	a	bcd
ab	1	1
cd	0	2

L'indice de Rand $R = 10/16$.

On notera que $1 - R = \frac{card(G_1 \Delta G_2)}{n^2}$ où $G_1 \Delta G_2$ est la différence symétrique des graphes induits par les deux partitions.

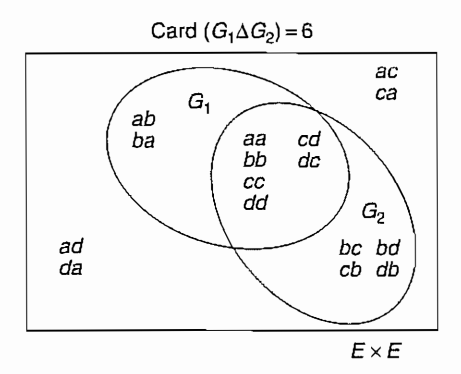


FIGURE 11.1

11.1.3 Aspects combinatoires de la classification

On pourrait penser que, muni d'un critère, la recherche de la meilleure partition soit chose facile : E étant fini, il suffirait de considérer toutes les partitions possibles (à nombre fixé de classes ou non) et de choisir celle qui optimise le critère de qualité choisi. Les résultats suivants montrent que cette tâche est insurmontable car le nombre de partitions devient vite astronomique : un calculateur pouvant traiter un million de partitions par seconde mettrait plus de 126 000 années pour étudier toutes les partitions d'un ensemble de 25 éléments !

Il faudra donc, dans la plupart des cas, se contenter de solutions approchées.

11.1.3.1 Nombre de partitions en k classes de n éléments

Notons $P_{n,k}$, ce nombre appelé nombre de Stirling de deuxième espèce. On a les résultats triviaux suivants : $P_{n,1} = P_{n,n} = 1$; $P_{n,n-1} = \frac{n(n-1)}{2}$. Le nombre de dichotomies

possibles est : $P_{n,2} = 2^{n-1} - 1$. En effet, il y a 2^n parties de E, donc $\frac{2^n}{2}$ partitions de E ou couples de parties complémentaires, mais parmi elles il y a la partition $\{E, \varnothing\}$ à éliminer.

Les nombres $P_{n,k}$ satisfont à l'équation de récurrence suivante qui permet de les calculer de proche en proche : (tableau 11.1)

$$P_{n,k} = P_{n-1,k-1} + kP_{n-1,k}$$

Démonstration : Soit une partition de E en k classes et soit un élément e de E : de deux choses l'une, ou bien e est seul dans sa classe, ou il ne l'est pas : si e est seul dans sa classe il y a $P_{n-1,k-1}$ partitions de cette sorte ; si e n'est pas seul dans sa classe c'est que $E - \{e\}$ est partitionné aussi en k classes et il y a $P_{n-1,k}$ manières de le faire et e peut se trouver alors dans l'une quelconque de ces k classes soit $kP_{n-1,k}$ possibilités. ■

On peut montrer que :

$$P_{n,k} = \frac{1}{k!}\sum_{i=1}^{k} C_k^i(-1)^{k-i}\, i^n$$

et donc si $n \to \infty$ $P_{n,k} \sim \frac{k^n}{k!}$.

11.1.3.2 Nombre total de partitions P_n (nombre de Bell)

On a :

$$P_n = \sum_{k=1}^{k=n} P_{n,k}$$

On peut aussi obtenir une formule de récurrence sur les P_n.

Considérons, comme précédemment, un élément e : pour une partition donnée de E, e se trouve dans une classe ; si cette classe a un élément, il y a P_{n-1} partitions de E laissant e seul dans une classe ; si cette classe a deux éléments, il y a C_{n-1}^1 manières de choisir le compagnon de e dans sa classe et P_{n-2} manières de constituer les autres classes ; si cette classe a k éléments, il y a C_{n-1}^{k-1} manières de choisir les compagnons de e et P_{n-k} manières de constituer les autres classes d'où :

$$P_n = P_{n-1} + C_{n-1}^1 P_{n-2} + C_{n-1}^2 P_{n-3} + \cdots + C_{n-1}^{k-1} P_{n-k} + \cdots + C_{n-1}^{n-2} P_1 + 1$$

Si l'on pose par convention $P_0 = 1$, on a la formule :

$$P_n = P_0 + (n-1)P_1 + C_{n-1}^2 P_2 + \cdots + C_{n-1}^k P_k + \cdots + P_{n-1}$$

On démontre que $P_n = \frac{1}{e}\sum_{k=1}^{\infty}\frac{k^n}{k!}$

TABLEAU 11.1

TABLE DES $P_{n,k}$

n \ k	1	2	3	4	5	6	7	8	9	10	11	P_n
1	1											1
2	1	1										2
3	1	3	1									5
4	1	7	6	1								15
5	1	15	25	10	1							52
6	1	31	90	65	15	1						203
7	1	63	301	350	140	21	1					877
8	1	127	966	1 701	1 050	266	28	1				4 140
9	1	255	3 025	7 770	6 951	2 646	462	36	1			21 147
10	1	511	9 330	34 105	42 525	22 827	5 880	750	45	1		115 975
11	1	1 023	28 501	145 750	246 730	179 487	63 987	11 880	1 155	55	1	678 970
12	1	2 047	86 526	611 501	1 379 400	1 323 652	627 396	159 027	22 275	1 705	66	4 213 597

11.1.4 Sur l'existence et la caractérisation des classes d'un ensemble

La définition de classes « naturelles » pose d'épineux problèmes. Si dans certaines situations simples comme celle de la figure 11.2 on voit clairement de quoi il s'agit, il est loin d'en être ainsi la plupart du temps et il faut bien admettre que l'on ne peut donner de définition claire des classes *a priori*. D'où la difficulté de valider des méthodes de classification en essayant de reconnaître des classes préexistantes.

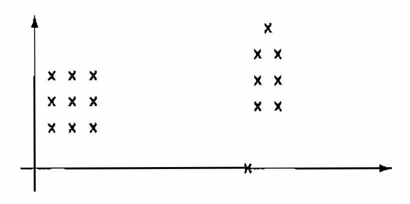


FIGURE 11.2

Bien souvent, les classes ne seront que ce qu'a produit un algorithme de classification.

Sur le plan pratique, la détermination du nombre « réel » de classes n'admet pas de solution satisfaisante.

Notons enfin qu'il ne suffit pas de produire des classes : il faut encore les interpréter et utiliser alors l'ensemble des informations disponibles et pas seulement les distances deux à deux.

11.2 LES MÉTHODES DE PARTITIONNEMENT

11.2.1 Les méthodes du type « nuées dynamiques » ou *k-means*

Ces méthodes permettent de traiter rapidement des ensembles d'effectif assez élevé en optimisant localement un critère de type inertie. On supposera que les individus sont des points de $\mathbb{R}^p$ muni d'une distance euclidienne.

11.2.1.1 Inertie interclasse et inertie intraclasse

Étant donné une partition en k groupes d'un nuage de n points, on définira les quantités suivantes : $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ centres de gravité des k groupes et $I_1, I_2, \ldots, I_k$ inerties des k groupes. On rappelle que l'inertie est la moyenne des carrés des distances au centre de gravité.

L'inertie totale I des n points autour du centre de gravité global $\mathbf{g}$ est alors égal à la somme de deux termes (théorème de König-Huyghens) :

$$I = I_B + I_W$$

où I_W est l'inertie intraclasse $I_W = \Sigma P_i I_i$, P_i étant le poids de la classe i et I_B l'inertie interclasse ou inertie du nuage des k centres de gravité : $I_B = \Sigma P_i d^2(\mathbf{g}_i, \mathbf{g})$.

Un critère usuel de classification consiste à chercher la partition telle que I_W soit minimal pour avoir en moyenne des classes bien homogènes, ce qui revient à chercher le maximum de I_B.

Remarquons que ce critère ne s'applique qu'à nombre de classes fixé : si k n'était pas fixé la solution serait la partition triviale en n classes (un individu = une classe) qui annule I_W.

11.2.1.2 La méthode des centres mobiles

Due à Forgy, elle consiste à partir de k points pris parmi E (en général tirés au hasard) ; ces k points définissent une partition de l'espace, donc une partition de E en k classes $E_{\mathbf{c}_1}, E_{\mathbf{c}_2}, \ldots, E_{\mathbf{c}_k}$. La partition de $\mathbb{R}^p$ associée à k centres $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$ est un ensemble de domaines polyédraux convexes déterminé par les hyperplans médiateurs des centres. $E_{\mathbf{c}_i}$ est la classe constituée par l'ensemble des points de E plus proches de $\mathbf{c}_i$ que de tout autre centre (fig. 11.3).

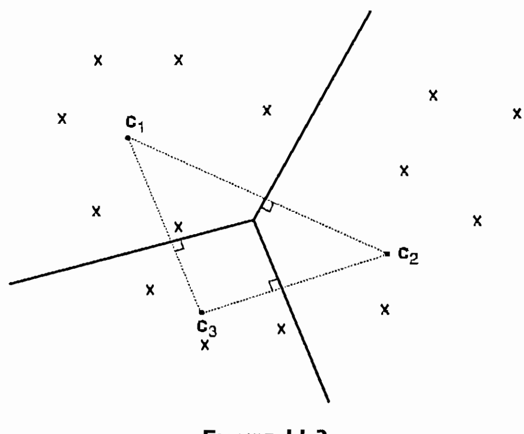


FIGURE 11.3

On remplace alors les k points pris au hasard par les k centres de gravité de ces classes et on recommence : l'algorithme converge rapidement vers un optimum local car le passage d'un centre arbitraire $\mathbf{c}_i$ à un centre de gravité diminue nécessairement la variance interne des classes.

Soit $E_{\mathbf{g}_i}$ la classe obtenue en remplaçant $\mathbf{c}_i$ par $\mathbf{g}_i$ centre de gravité de $E_{\mathbf{c}_i}$. Il suffit de montrez que :

$$\frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j\in E_{\mathbf{c}_i}} d(\mathbf{j},\mathbf{g}_i)^2\right) \geq \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{\mathbf{k}\in E_{\mathbf{g}_i}} d(\mathbf{k},\mathbf{g}_i)^2\right)$$

car, d'après le théorème de König-Huyghens, $\mathbf{g}_i$ n'étant pas le centre de gravité de $E_{\mathbf{g}_i}$, le membre de droite sera supérieur à la variance intraclasse de la partition $E_{\mathbf{g}_i}$.

Or, si l'on considère un point quelconque, il figurera dans le membre de droite avec son carré de distance au $\mathbf{g}_i$ qui sera le plus proche de lui par construction des $E_{\mathbf{g}_i}$, tandis que dans le membre de gauche il figurera avec sa distance à un $\mathbf{g}_i$ qui ne sera pas forcément le plus proche de lui, mais qui sera seulement son centre de gravité dans la partition $E_{\mathbf{c}_i}$.

Ceci démontre donc le résultat annoncé : le nuage étant fini (l'ensemble de toutes les partitions possibles aussi), l'algorithme converge car la variance intraclasse ne peut que diminuer ; elle atteindra donc son minimum accessible à partir du système initial de centres $\mathbf{c}_i$ en un nombre fini d'itérations, le théorème de Huyghens indiquant que cette décroissance est stricte si $\mathbf{g}_i$ n'est pas confondu avec $\mathbf{c}_i$.

L'expérience montre que le nombre d'itérations nécessaires est très faible. Si au cours d'une itération une classe se vide, il est possible de tirer au hasard un nouveau centre.

La figure 11.4 montre le déroulement de l'algorithme sur un cas d'école : on voit qu'en deux itérations on a « reconnu » la structure existante.

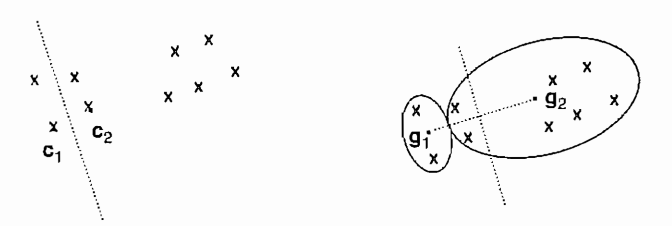


FIGURE 11.4

Cette méthode peut s'appliquer à des données qualitatives à condition de travailler sur des coordonnées factorielles. On a ainsi effectué une partition en quatre classes de l'ensemble des 27 chiens étudiés au chapitre précédent en utilisant les coordonnées issues de l'ACM du tableau disjonctif (on utilise ici les 10 facteurs). Quatre centres ont été tirés au hasard (les individus 11, 7, 18, 25) et on a abouti en moins de 10 itérations à la partition suivante :

Classe 1 : individus n° 2, 5, 7, 8, 17, 22, 26.
Classe 2 : individus n° 4, 9, 11, 14, 19.
Classe 3 : individus n° 6, 21, 24, 27.
Classe 4 : individus n° 1, 3, 10, 12, 13, 15, 16, 18, 20, 23, 25.

On a : inertie interclasse = 0.93665, inertie intraclasse = 0.73001, l'inertie totale valant $1.66667 = \left(\frac{1}{p}\sum m_i\right) - 1$.

Dans la méthode précedente, on attend que tous les individus aient été affectés à une classe pour recalculer les centres de gravité. La variante de Mac Queen procède différemment : les centres sont recalculés après l'affectation de chaque point.

La méthode des nuées dynamiques, proposée par E. Diday, est une extension de la précédente. Elle en diffère notamment par les traits suivants : au lieu de représenter une classe uniquement par son centre de gravité, on la caractérise par un « noyau ». Ce noyau peut être un ensemble de q points (les plus centraux), un axe principal ou un plan principal, etc.

Il faut donc disposer formellement d'une fonction de représentation qui, à un ensemble de points, associe son noyau.

Il faut ensuite disposer d'un algorithme de réaffectation des points aux noyaux. On procède alors par alternance des deux phases : affectation, représentation jusqu'à convergence du critère choisi. La méthode des nuées dynamiques ne se limite pas au cas de distances euclidiennes.

Comme la partition finale peut dépendre du tirage des noyaux de départ (problème d'optimum local), on recommence alors toute l'opération avec s autres tirages. On appelle « formes fortes » ou « groupements stables » les ensembles d'éléments ayant toujours été regroupés lors de la partition finale pour les s passages de l'algorithme.

11.2.2 La méthode de Condorcet

Considérons un ensemble de n individus décrits par p variables qualitatives à $m_1, m_2, \ldots, m_p$ modalités respectivement : on a p partitions différentes du même ensemble.

La recherche d'une nouvelle partition revient donc à rechercher un compromis entre ces p partitions initiales.

Soit $\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^p$ les tableaux des comparaisons par paires associés à $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$, et $\mathbf{C} = \sum_k \mathbf{C}^k$.

c_{ij} est le nombre de fois parmi p où les objets i et j ont été mis dans une même classe.

Soit $\mathbf{C}' = 2\mathbf{C} - p$. On a alors $c'_{ij} > 0$ si i et j sont dans une même classe pour une majorité de variables $\mathcal{X}_k$, $c'_{ij} < 0$ si il y a une majorité de variables où i et j sont dans des classes différentes ; $c'_{ij} = 0$ s'il y a autant de variables pour lesquelles i et j sont séparés que de variables pour lesquelles i et j sont réunis.

Un critère naturel pour former une partition « centrale », compromis entre les p partitions, consiste alors à mettre i et j dans une même classe à chaque fois que c'_{ij} est positif et à les séparer à chaque fois que c'_{ij} est négatif. Malheureusement, ce critère ne fournit pas nécessairement une partition : il peut y avoir non transitivité de la règle majoritaire. C'est le paradoxe de Poincaré : ce n'est pas parce qu'il y a une majorité pour réunir i et j, j et k qu'il y a une majorité pour réunir i et k.

Il faut donc imposer les contraintes des relations d'équivalence ce qui revient à chercher la partition satisfaisant au mieux la majorité des partitions initiales.

Si **Y** est le tableau des comparaisons par paires de la partition cherchée, on a donc à résoudre le problème suivant :

$$\max \sum_i \sum_j c'_{ij} y_{ij}$$

avec : $\begin{cases} y_{ij} + y_{ji} \\ y_{ij} + y_{jk} - y_{ik} \leq 1 \\ y_{ij} = 0 \text{ ou } 1 \end{cases}$

C'est un problème de programmation linéaire bivalente dont on peut trouver une solution exacte (pas forcément unique) si n est faible, ou une solution approchée si n n'est pas trop élevé en utilisant des heuristiques (voir l'ouvrage de Marcotorchino et Michaud cité en référence).

Il y a, en effet, de l'ordre de n^2 inconnues $\left(\dfrac{n(n-1)}{2} \text{exactement}\right)$ et de l'ordre de n^3 contraintes.

On aura remarqué que le nombre de classes n'a pas à être imposé, il fait partie de la solution.

La distance de la différence symétrique entre les deux partitions associées aux tableaux $\mathbf{C}^k$ et **Y** vaut :

$$\sum_i \sum_j |y_{ij} - c^k_{ij}| = \sum_i \sum_j (y_{ij} - c^k_{ij})^2 = \sum_i \sum_j c^k_{ij} - \sum_i \sum_j c'^k_{ij} y_{ij} = d(\mathbf{C}^k, \mathbf{Y})$$

La partition cherchée est donc celle qui est à distance moyenne minimale des partitions initiales puisque :

$$\min_Y \sum_k d(\mathbf{Y}, \mathbf{C}^k) = \min \left[\sum_i \sum_j c_{ij} - \sum_i \sum_j c'_{ij} y_{ij} \right]$$

ce qui revient à chercher $\max \sum_i \sum_j c'_{ij} y_{ij}$.

D'après le paragraphe 11.1.2, la partition optimale est donc celle qui maximise la somme des indices de Rand avec chacune des partitions associées aux p variables qualitatives. On retrouve ici une propriété d'***association maximale*** :

$$\boxed{\max \sum_k \Phi\,(\mathcal{Y}, \mathcal{X}^k)}$$

où Φ est un critère d'association entre variables qualitatives.

Lorsque Φ n'est pas l'indice de Rand, il faut en général fixer le nombre de classes de $\mathcal{Y}$. En l'exprimant en termes de comparaison par paires, c'est-à-dire en explicitant la mesure d'association Φ en fonction des tableaux Y et C^k, on peut se ramener à un problème de programmation linéaire dans certains cas.

Ainsi l'indice d'association de Belson entre deux variables qualitatives défini par :

$$\Phi(\mathcal{Y}, \mathcal{X}^k) = \sum_u \sum_v \left(n_{uv} - \frac{n_{u.}\, n_{.v}}{n} \right)^2$$

où n_{uv} est le terme général du tableau de contingence croisant $\mathcal{Y}$ et $\mathcal{X}^k$ s'écrit avec les tableaux $\mathbf{Y}$ et $\mathbf{C}^k$:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(c_{ij}^k - \frac{c_{i.}^k + c_{.j}^k}{n} + \frac{c_{..}^k}{n^2}\right)y_{ij} = \sum_i\sum_j a_{ij}^k y_{ij}$$

11.3 MÉTHODES HIÉRARCHIQUES

Elles consistent en un ensemble de partitions de E en classes de moins en moins fines obtenues par regroupements successifs de parties. Une classification hiérarchique se représente par un ***dendrogramme*** ou arbre de classification (fig. 11.5) :

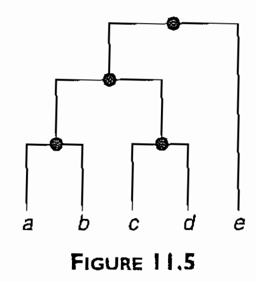


FIGURE 11.5

Cet arbre est obtenu de manière ascendante dans la plupart des méthodes : on regroupe d'abord les deux individus les plus proches qui forment un « noeud », il ne reste plus que $n - 1$ objets et on itère le processus jusqu'à regroupement complet. Un des problèmes consiste à définir une mesure de dissimilarité entre parties. Les méthodes descendantes, ou algorithmes divisifs, sont plus rarement utilisées.

11.3.1 Aspect formel

11.3.1.1 Hiérarchie de parties d'un ensemble E

Une famille H de parties de E est une hiérarchie si :

a) E et les parties à un élément appartiennent à H.
b) $\forall A, B \in H \quad A \cap B \in \{A, B, \varnothing\}$. En d'autres termes, deux classes sont soit disjointes, soit contenues l'une dans l'autre.
c) Toute classe est la réunion des classes qui sont incluses en elle.

A toute hiérarchie correspond un arbre de classification :

Exemple : $H = \{\varnothing, a, b, c, d, e, f\,; ab\,; abc, de, def, abcdef\}$ (fig. 11.6)

Une partition de E compatible avec H est une partition dont les classes sont des éléments de H. D'une manière imagée, c'est une partition obtenue en coupant l'arbre selon une horizontale et en recueillant les morceaux.

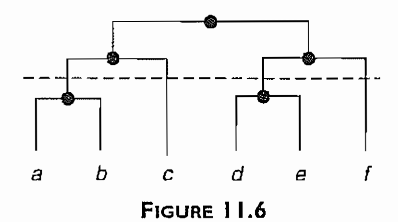


FIGURE 11.6

Lorsque l'on peut dire qu'un élément ou une partie A est reliée à B avant que C ne soit reliée à D, autrement dit s'il existe une relation de préordre totale compatible avec la relation d'ordre naturelle par inclusion, on dit qu'on a une hiérarchie ***stratifiée***.

Une hiérarchie est ***indicée*** s'il existe une application i de H dans $\mathbb{R}^+$ croissante, c'est-à-dire telle que si $A \subset B$: $i(A) \leq i(B)$. A toute hiérarchie indicée correspond une hiérarchie stratifiée. Les indices sont appelés niveaux d'agrégation : $i(A)$ est le niveau auquel on trouve agrégés pour la première fois tous les constituants de A. Ainsi, dans la figure 11.7 on a $i(c, d) = 1$ et $i(a, b, c) = 0.5$.

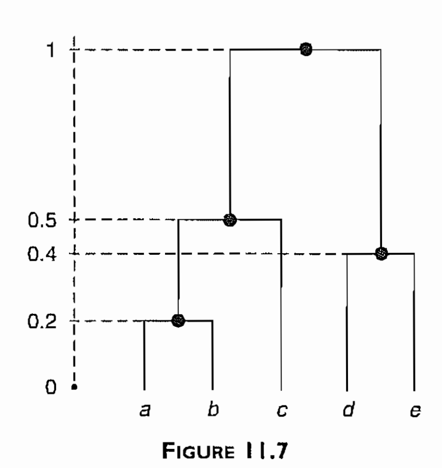


FIGURE 11.7

Les niveaux d'agrégation sont pris égaux, en général, à l'indice de dissimilarité des deux parties constituant la réunion $i(a, b, c) = \delta((a, b), c)$.

Le problème se pose alors de savoir si la hiérarchie peut présenter ou non des inversions : si a, b sont réunis avant c, d dans l'algorithme, on doit avoir $i(a, b) < i(c, d)$ sinon il y a inversion.

11.3.1.2 Distances ultramétriques

A toute hiérarchie indicée H correspond un indice de distance entre éléments de H : $d(A, B)$ est le niveau d'agrégation de A et de B, c'est-à-dire l'indice de la plus petite partie de H contenant à la fois A et B.

Cette distance possède la propriété suivante, dite propriété ultramétrique :

$$d(a, b) \leq \sup \{d(a, c) ; d(b, c)\} \quad \forall a, b, c$$

En effet, de deux choses l'une, quand « a » a été réuni à « b » pour la première fois :

- ou bien c n'est pas encore réuni à a (ni à b par conséquent), il sera donc réuni plus tard, donc $d(a, c)$ qui est égal à $d(b, c)$, puisque a et b sont maintenant réunis, est supérieur à $d(a, b)$;
- ou bien c est déjà réuni à a ou b, supposons à a pour fixer les idées, avant que a ne soit réuni à b. Donc $d(a, c) < d(a, b)$. Mais alors $d(b, c) = d(a, b)$, car c est réuni à b en même temps que b l'est à a. Ce qui démontre la relation ultramétrique.

Réciproquement, à toute ultramétrique correspond une hiérarchie indicée ; la recherche d'une classification hiérarchique est donc équivalente à celle d'une ultramétrique ; le problème clé de la classification est donc le suivant : connaissant une métrique sur E, en déduire une ultramétrique aussi proche que possible de la métrique de départ.

Les propriétés suivantes de géométrie ultramétrique précisent le lien avec les hiérarchies indicées :

- En géométrie ultramétrique, tout triangle est soit isocèle pointu (la base est inférieure à la longueur commune des deux autres côtés), soit équilatéral.

En effet :

$$d(a, c) \leq \sup \{d(a, c) ; d(b, c)\}$$
$$d(a, c) \leq \sup \{d(a, b) ; d(b, c)\}$$
$$d(b, c) \leq \sup \{d(a, b) ; d(b, c)\}$$

Supposons par exemple $d(a, b) > d(a, c) > d(b, c)$. Cette hypothèse est absurde car une des trois relations ultramétriques n'est plus vérifiée. Il faut donc que deux côtés soient égaux et on voit aisément que ce sont forcément les deux plus grands qui le sont.

- En géométrie ultramétrique, tout point d'une boule est centre de cette boule.

En effet, soit B la boule ensemble des points dont la distance à un centre a est inférieur à r : $B(a, r) = \{x \mid d(a, x) \leq r\}$.

Soient x et y deux points $\in B$: $d(x, y) \leq \sup \{d(x, a) ; d(a, y)\}$.

On en déduit que, si deux boules ont une intersection non vide, l'une est nécessairement incluse dans l'autre puisqu'elles sont concentriques. On retrouve bien ici la propriété d'inclusion des parties d'une hiérarchie.

11.3.2 Stratégies d'agrégation sur dissimilarités

On suppose ici que l'on connaît un indice de dissimilarité d. Différentes solutions existent qui correspondent à des choix différents de la dissimilarité entre parties de E, appelés stratégies. Le problème est en effet de définir la dissimilarité entre la réunion de deux éléments et un troisième : $d((a, b) ; c)$. A chaque solution correspond une ultramétrique différente.

11.3.2.1 Le saut minimum

Cette méthode (connue sous le nom de *single linkage* en anglais) consiste à écrire que : $d((a, b) ; c) = \inf(d(a, c) ; d(b, c))$. La distance entre parties est donc la plus petite distance (fig. 11.8) entre éléments des deux parties.

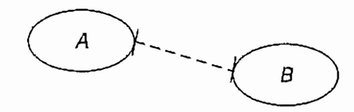


FIGURE 11.8

Cette stratégie conduit à une ultramétrique particulière : la « sous-dominante » qui est parmi les ultramétriques inférieures à d ($\delta(i, j) \leq d(i, j)$) la plus élevée ($\delta(i, j)$ maximum).

En effet, la construction de l'arbre aboutit à une suite de partitions P_h emboîtées, conduisant chacune à une dissimilarité d_h entre parties.

Nous allons montrer que d_h, qui est forcément inférieure à δ, est supérieure à toute ultramétrique inférieure à δ : comme, à la limite, d_h devient ultramétrique, c'est que l'on a obtenu l'ultramétrique inférieure maximale.

Montrons ceci par récurrence ; si c'est vrai pour d_{h-1}, montrons que c'est encore vrai pour d_h. Il suffit d'examiner les couples pour lesquels $d_h(u, i) \neq d_{h-1}(u, i)$. Ceci n'est possible que si u (ou i) vient d'être agrégé.

Nous sommes donc dans la situation suivante où, au pas h, on vient d'agréger i à i' (fig. 11.9).

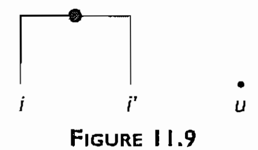


FIGURE 11.9

Si l'on a agrégé i à i', c'est qu'ils étaient les plus proches avec la dissimilarité d_{h-1} et l'on a alors :

$$d_h(i, i') = d_{h-1}(i, i') \leq \inf \{d_{h-1}(u, i), d_{h-1}(u, i')\}$$

D'autre part, on a précisément $d_h(u, i) = d_h(u, i') = \inf \{d_{h-1}(u, i), d_{h-1}(u, i)\}$ par hypothèse.

Soit « d » une ultramétrique inférieure ou égale à δ, donc à d_{h-1} (récurrence) $d(u, i) \leq d_{h-1}(u, i)$.

Comme d est ultramétrique, $d(u, i) \leq \sup [d(i, i') ; d(u, i')\}$ donc :

$$d(u, i) \leq \sup \{d_{h-1}(i, i') ; d_{h-1}(u, i')\}$$

Comme $d_{h-1}(i, i') \leq \inf [d_{h-1}(u, i), d_{h-1}(u, i')]$. On a $d(u, i) \leq d_{h-1}(u, i')$.

On a donc à la fois $d(u, i) \leq d_{h-1}(u, i')$, $d(u, i) \leq d_{h-1}(u, i)$ et :

$$d_h(u, i) = \inf \{d_{h-1}(u, i') ; d_{h-1}(u, i')\}$$

C'est donc que $d(u, i) \leq d_h(u, i)$.

Une autre méthode pour aboutir à l'ultramétrique inférieure maximale, due à M. Roux, consiste à passer en revue tous les triangles possibles faits avec les points de E et à les rendre isocèles pointus (on remplace la longueur du plus grand côté par celle du côté médian), de manière à obtenir directement l'ultramétrique inférieure maximale. On passe en revue tous les triangles jusqu'à ce qu'on ne puisse plus rien modifier ; le reste ensuite á tracer l'arbre.

11.3.2.2 Le diamètre et autres stratégies

On prend ici comme distance entre parties la plus grande distance :

$$d((a, b) ; c) = \sup(d(a, c), d(b, c))$$

On aboutit alors à une des ultramétriques supérieures minimales, contrairement au cas précédent où la sous-dominante est unique. Il n'existe pas en effet une seule ultramétrique minimale parmi les ultramétriques supérieures à d ; on montre même qu'il en existe $(n - 1)!$ dans le cas où toutes les valeurs de la dissimilarité sont différentes.

De nombreuses autres méthodes de calcul de distances entre parties ont été proposées (moyenne des distances, etc.) toutes sont des cas particuliers de la formule de Lance et Williams généralisée par Jambu :

$$d((a, b) ; c) = a_1 d(a, c) + a_2 d(b, c) + a_3 d(a, b) + a_4 i(a) + a_5 i(b) + a_6 i(c) + a_7 |d(a, b) - d(b, c)|$$

Pour qu'il n'y ait pas d'inversion, il faut que les coefficients vérifient :

$$\begin{cases} a_1 + a_2 + a_3 \geq 1 \\ a_1, a_2, a_3, a_4, a_5, a_6 \geq 0 \\ a_7 \geq -\min(a_1 ; a_2) \end{cases}$$

Ainsi la méthode du saut minimal consiste à prendre :

$$a_1 = a_2 = 1/2, a_3 = a_4 = a_5 = a_6 = 0, a_7 = 1/2$$

11.3.3 La méthode de Ward pour distances euclidiennes

Si l'on peut considérer E comme un nuage d'un espace $\mathbb{R}^p$, on agrège les individus qui font le moins varier l'inertie intraclasse. En d'autres termes, on cherche à obtenir à chaque pas un minimum local de l'inertie intraclasse ou un maximum de l'inertie interclasse.

L'indice de dissimilarité entre deux classes (ou niveau d'agrégation de ces deux classes) est alors égal à la perte d'inertie interclasse résultant de leur regroupement.

Calculons cette perte d'inertie.

Soit $\mathbf{g}_A$ et $\mathbf{g}_B$ les centres de gravité de deux classes et $\mathbf{g}_{AB}$ le centre de gravité de leur réunion. On a :

$$\mathbf{g}_{AB} = \frac{p_A \mathbf{g}_A + p_B \mathbf{g}_B}{p_A + p_B}$$

où p_A et p_B sont les poids des deux classes (fig. 11.10).

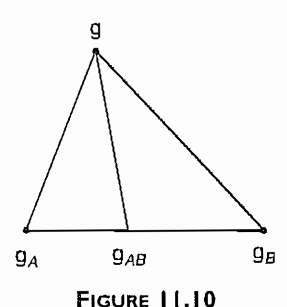


FIGURE 11.10

L'inertie interclasse étant la moyenne des carrés des distances des centres de classe au centre de gravité total, la variation d'inertie est égale à :

$$p_A d^2(\mathbf{g}_A, \mathbf{g}) + p_B d^2(\mathbf{g}_B, \mathbf{g}) - (p_A + p_B) d^2(\mathbf{g}_{AB}, \mathbf{g})$$

Un calcul élémentaire montre que cette variation vaut $\dfrac{p_A p_B}{p_A + p_B} d^2(\mathbf{g}_A, \mathbf{g}_B)$ (qui est donc positive).

Si l'on pose :
$$\delta(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(\mathbf{g}_A, \mathbf{g}_B)$$

Cette méthode rentre dans le cadre de la formule de Lance et Williams généralisée car :

$$\delta((A, B)\,;\, C) = \frac{(p_A + p_C)\delta(A, C) + (p_B + p_C)\delta(B, C) - p_C\delta(A, B)}{p_A + p_B + p_C}$$

on peut donc utiliser l'algorithme général.

On notera que la somme des niveaux d'agrégation des différents nœuds de l'arbre est égale à l'inertie totale du nuage puisque la somme des pertes d'inertie est égale à l'inertie totale.

Cette méthode est donc complémentaire de l'analyse en composantes principales et repose sur un critère d'optimisation assez naturel. Elle constitue à notre avis la méthode de classification hiérarchique de référence sur données euclidiennes. Il ne faut pas oublier cependant que le choix de la métrique dans l'espace des individus conditionne également les résultats.

11.3.4 Classification de données qualitatives

Lorsque les n individus à classer sont décrits par des variables qualitatives, divers cas se présentent. Pour les données de présence-absence, on utilisera un des indices de dissimilarité présentés au paragraphe 11.1.1.2.

Pour des données du type p variables qualitatives à $m_1, m_2, \ldots, m_p$ modalités, on utilisera la représentation disjonctive complète :

$$\left[\begin{array}{c|c|c} 0100 & \ldots & 001 \\ 1000 & \ldots & 001 \end{array}\right] = \mathbf{X}$$

La distance du χ^2 entre lignes du tableau possède alors des propriétés intéressantes :

$$d^2(i, i') = \sum_j \frac{n}{n_{.j}}\left(\frac{x_{ij} - x_{i'j}}{p}\right)^2$$

L'indice de similarité associé à d^2 est alors le produit scalaire du χ^2 : $\sum_j \frac{n}{n_{.j}}(x_{ij}\, x_{i'j})$ x_{ij} étant égal à 0 ou 1. On voit que la similarité dépend non seulement du nombre de modalités possédées en commun par i et i' mais de leur fréquence, ce qui revient à dire que deux individus qui ont en commun une modalité rare sont plus proches que deux individus ayant en commun une modalité fréquente : cette propriété semble assez naturelle.

On utilisera alors la méthode de Ward (puisque la distance χ^2 est euclidienne) sur le tableau des distances.

Une autre solution consiste à effectuer une classification hiérarchique sur le tableau des coordonnées des n individus après analyse des correspondances multiples de **X**. Il faut prendre garde ici que ces deux approches ne seront équivalentes qu'à la condition d'utiliser **tous** les facteurs de l'ACM (soit $\sum_{i=1}^{p} m_i - p$). En effet, une classification effectuée sur un trop petit nombre de facteurs peut être fallacieuse car elle peut laisser de côté certaines particularités du nuage de points. Par ailleurs, il ne faut pas oublier de conserver la normalisation à $\sqrt{\lambda}$ de chaque axe car ceux-ci ont des importances différentes. Ces remarques sont valables également pour des classifications effectuées sur des composantes principales.

La classification hiérarchique des lignes ou des colonnes d'un tableau de contingence s'effectuera avec la méthode de Ward et la distance du χ^2 entre lignes (ou entre colonnes). Cette méthode revient à regrouper les catégories d'une variable qualitative de la façon suivante : à chaque étape, on réunit les deux catégories (en sommant les effectifs) qui font diminuer le moins possible le φ^2 puisque l'inertie totale est ici $\chi^2/n = \varphi^2$.

11.3.5 Considérations algorithmiques

L'algorithme général consiste à balayer à chaque étape un tableau de $\frac{n(n-1)}{2}$ distances ou dissimilarités afin d'en rechercher l'élément de valeur minimale, à réunir les deux individus correspondant, à mettre à jour les distances après cette réunion et à recommencer avec $n - 1$ objets au lieu de n.

La complexité d'un tel algorithme est en n^3 (ordre du nombre d'opérations à effectuer) et on atteint rapidement les limites d'un ordinateur même puissant pour quelques centaines d'observations.

Diverses techniques ont été proposées pour accélérer les opérations et pouvoir traiter des ensembles plus vastes d'individus.

La méthode des voisinages réductibles (M. Bruynhooghe) consiste à n'effectuer les comparaisons de distances que pour celles qui sont inférieures à un seuil fixé. Il faut ensuite réactualiser ce seuil au fur et à mesure que la classification s'effectue.

La méthode des voisins réciproques (Mac Quitty et J. P. Benzecri) consiste à réunir simultanément plusieurs paires d'individus (les voisins réciproques) à chaque lecture du tableau

des distances, la complexité de l'algorithme devient alors en n^2. La recherche des voisins réciproques s'effectue alors en chaîne : on part d'un objet quelconque et on cherche son plus proche voisin, puis le plus proche voisin de celui-ci, etc., jusqu'à aboutir à un élément dont le plus proche voisin est son prédécesseur dans la liste. On réunit ces deux éléments et on recommence à partir du nœud créé ou de l'avant-dernier élément de la liste jusqu'à création de tous les nœuds.

11.4 MÉTHODES MIXTES POUR GRANDS ENSEMBLES

La détermination du nombre de classes est relativement aisée en classification hiérarchique en étudiant le dendrogramme et en s'aidant de l'histogramme des indices de niveau. La coupure de l'arbre en k classes ne fournit cependant pas la partition optimale en k classes de l'ensemble en raison de la contrainte d'emboîtement des partitions issues d'une hiérarchie. Mais cette coupure fournit une excellente initialisation pour un algorithme de partitionnement de type nuées dynamiques. De cette façon on peut résoudre pratiquement le problème épineux du choix du nombre de classes d'une partition. Cependant les méthodes de classification hiérarchique ne sont pas utilisables lorsque le nombre d'individus à classer est trop élevé (supérieur à plusieurs milliers), alors que les méthodes de partitionnement ne connaissent pas ce genre de limites et sont très rapides.

Le principe des méthodes mixtes, également apelées hybrides, tire parti des avantages des deux techniques. Concrètement, on procéde de la façon suivante en trois étapes :

1. Recherche d'une partition en un grand nombre K de classes (par exemple 100) avec une méthode de type nuées dynamiques
2. Regroupement hiérarchique des K classes à partir de leurs centres de gravité et détermination d'une coupure en k classes
3. Consolidation : amélioration de la partition en k classes par une méthode de type nuées dynamiques

11.5 CLASSIFICATION DE VARIABLES

La plupart des méthodes exposées précédemment ont été conçues pour classer des individus.

Lorsque l'on veut faire des regroupements de variables, il convient de prendre certaines précautions car la notion de distance entre deux variables pose souvent de délicats problèmes dus à la nature des variables.

11.5.1 Variables numériques

Pour des variables numériques, le coefficient de corrélation linéaire constitue l'indice naturel et $1 - r$ est alors un indice de dissimilarité qui est en plus une distance euclidienne.

On peut alors utiliser la méthode hiérarchique de Ward et celle des nuées dynamiques puisque l'on dispose d'une distance euclidienne.

Une variante consiste à utiliser les coordonnées des variables sur des axes factoriels.

Mentionnons également la méthode divisive (ou descendante) disponible dans le logiciel SAS (procédure varclus) qui revient à déterminer les groupes de variables les plus unidimensionnels possible au sens où l'ACP de chaque groupe ne fournit qu'une seule

dimension : une seule valeur propre supérieure à 1. L'algorithme est sommairement le suivant : on part de l'ensemble des p variables et on effectue une ACP. Si il n'y a qu'une seule valeur propre supérieure à 1, on s'arrête. Sinon on classe les variables en deux groupes selon leurs proximités avec la premiere ou la deuxiéme composante principale. On recommence alors la procédure dans chaque groupe.

11.5.2 L'approche de Lerman et l'algorithme de la vraisemblance du lien

Pour des variables qualitatives, un problème vient du fait que les mesures de liaison ne sont comparables que pour des nombres égaux de catégories, ou du degré de liberté du couple.

I. C. Lerman a proposé de remplacer la valeur de l'indice de similarité entre variables de même nature (corrélation, χ^2, etc.) par la probabilité de trouver une valeur inférieure dans le cadre de l'hypothèse d'indépendance (appelée « absence de lien »). Ainsi, au lieu de prendre r, on prendra $P(R < r)$. L'avantage est incontestable pour les mesures de similarité entre variables qualitatives qui deviennent dès lors comparables indépendamment des nombres de catégories : un χ^2_4 égal à 4 correspond à une similarité de 0.6 alors qu'un χ^2_{10} égal à 5 correspond à une similarité de 0.12.

L'algorithme de la vraisemblance du lien (AVL) consiste alors à utiliser comme mesure de proximité entre deux groupes A et B de m et l variables respectivement, la probabilité associée à la plus grande valeur observée de l'indice probabiliste de similarité.

Soit: $$t_0 = \sup_{\substack{x \in A \\ y \in B}} s(x, y)$$

où $s(x, y) = P(R < r(x, y))$ par exemple.

Dans l'hypothèse d'absence de lien, on a :

$$P(\sup_{x \in A} s(x, y) < t) = t^m$$

(voir chapitre 12, paragr. 12.1.3.2), d'où :

$$P(\sup_{\substack{x \in A \\ x \in B}} s(x, y) < t) = (t^m)^l = t^{ml}$$

On prendra donc comme indice de dissimilarité entre A et B : t_0^{ml}.

On peut alors obtenir une classification hiérarchique des variables.

11.6 EXEMPLES

Reprenons ci-dessous les différents exemples déjà étudiés dans les chapitres précédents, pour montrer la complémentarité entre les méthodes factorielles et les méthodes de classification.

11.6.1 Données voitures

Les données étant euclidiennes, on utilisera tout d'abord la méthode de Ward sur données réduites.

Le tableau suivant donne l'historique des regroupements. On vérifie que la somme des indices de niveau (ou somme des pertes d'inertie) est égale à l'inertie totale. L'appellation « ainé » « benjamin » est sans signification et ne fait que désigner les deux éléments réunis. On constate des sauts importants après le nœud 34 quand on passe de 3 classes à deux classes. Une coupure de l'arbre en 3 classes est alors naturelle.

DESCRIPTION DES NOEUDS

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
19	12	7	2	2.00	0.01417	*
20	16	5	2	2.00	0.02432	*
21	6	3	2	2.00	0.03061	*
22	18	4	2	2.00	0.03581	**
23	17	14	2	2.00	0.04593	**
24	21	15	3	3.00	0.06556	**
25	8	22	3	3.00	0.07693	***
26	20	11	3	3.00	0.08478	***
27	2	19	3	3.00	0.11771	****
28	25	10	4	4.00	0.12485	****
29	23	27	5	5.00	0.17459	******
30	9	13	2	2.00	0.22307	*******
31	28	1	5	5.00	0.23849	*******
32	26	24	6	6.00	0.36099	***********
33	29	32	11	11.00	0.52497	****************
34	30	33	13	13.00	1.06604	*******************************
35	34	31	18	18.00	2.79117	***************************************

SOMME DES INDICES DE NIVEAU = 6.00000

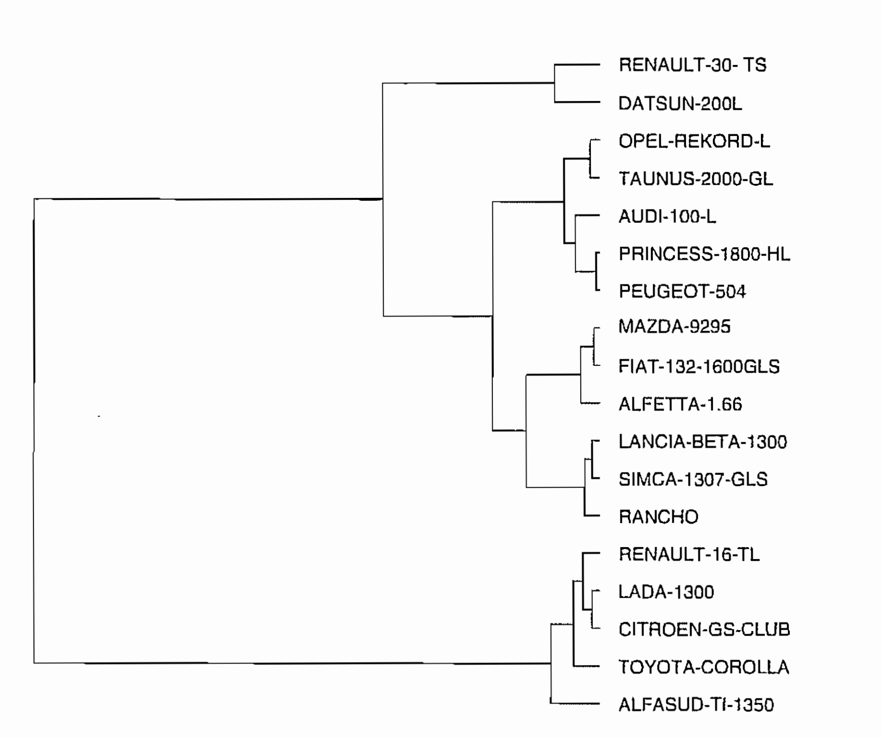


FIGURE 11.11

La coupure en 3 classes semble optimale car aucune amélioration n'est obtenue après passage d'une méthode de centres mobiles :

```
CONSOLIDATION DE LA PARTITION AUTOUR DES 3 CENTRES DE CLASSES, REALISEE PAR
10 ITERATIONS A CENTRES MOBILES ; PROGRESSION DE L'INERTIE INTER-CLASSES
```

ITERATION	I.TOTALE	I.INTER	QUOTIENT
0	6.00000	3.85720	0.64287
1	6.00000	3.85720	0.64287
2	6.00000	3.85720	0.64287

```
ARRET APRES L'ITERATION 2 L'ACCROISSEMENT DE L'INERTIE INTER-CLASSES PAR RAP-
PORT A L'ITERATION PRECEDENTE N'EST QUE DE 0.000 %.
```

La figure suivante donne dans le plan 1-2 la visualisation des 3 classes.

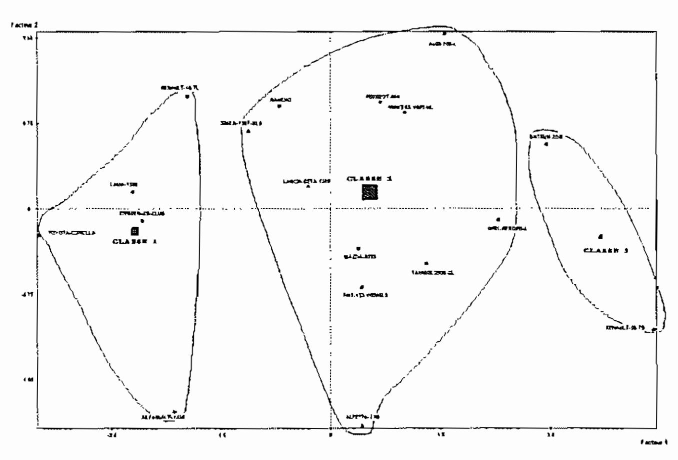


FIGURE 11.12

Ces 3 classes correspondent pour l'essentiel à la taille des individus

11.6.2 Vacances

Reprenons maintenant le tableau de contingence étudié en 9.3 avec une AFC.

La distance du khi-deux entre profils-lignes ou profils-colonnes étant une disatnce euclidienne, il est ici possible d'effectuer deux classifications, l'une sur les lignes, l'autre sur les colonnes du tableau de contingence .

11.6.2.1 Classification des professions

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
9	6	5	2	5027.00	0.00239	****
10	2	1	2	1375.00	0.00274	*****
11	4	10	3	5162.00	0.00473	********
12	9	8	3	5419.00	0.00587	**********
13	3	11	4	9511.00	0.01107	******************
14	12	13	7	14930.00	0.03125	**
15	7	14	8	18532.00	0.04930	**

SOMME DES INDICES DE NIVEAU = 0.10734

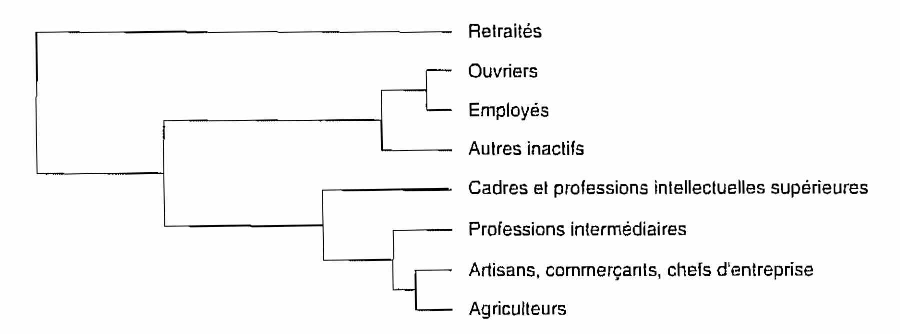


FIGURE 11.13

11.6.2.2 Classification des modes d'hébergement

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
10	2	8	2	3479.00	0.00025	*
11	4	9	2	6936.00	0.00276	*****
12	3	1	2	4364.00	0.00660	**********
13	7	6	2	1903.00	0.00674	***********
14	11	13	4	8839.00	0.00864	**************
15	10	5	3	5329.00	0.00969	***************
16	14	15	7	14168.00	0.01974	******************************
17	16	12	9	18532.00	0.05291	**

SOMME DES INDICES DE NIVEAU = 0.10734

On vérifie dans les deux cas que la somme des indices de niveau est bien égale au phi-deux de Pearson.

On constate que l'on pourrait regrouper aussi bien les lignes que les colonnes en 3 modalités.

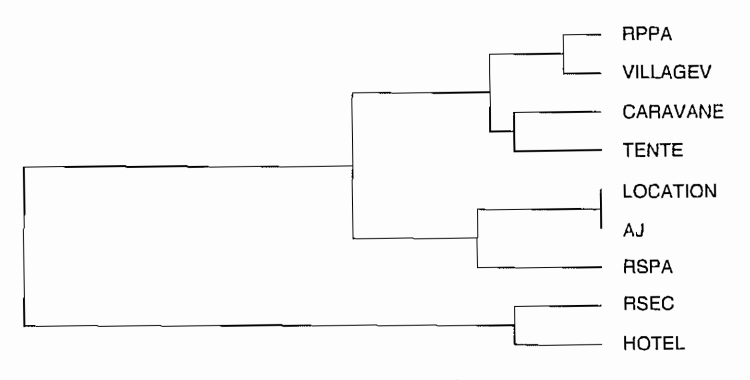


FIGURE 11.14

11.6.3 Races canines

Les données du chapitre 10 concernant 27 races canines ont été soumises à une classification ascendante hiérarchique selon la méthode de Ward sur les 10 composantes de l'analyse des correspondances multiples.

On trouve ci-après la liste de formation des nœuds et le dendrogramme (fig. 11.15) et tableau 11.2. Il est clair qu'une coupure est à effectuer au-dessus du nœud n° 50 (coude dans le diagramme des indices de niveau) et que l'on distingue nettement quatre classes homogènes, ces classes sont ici voisines de celles obtenues par la méthode des centres mobiles.

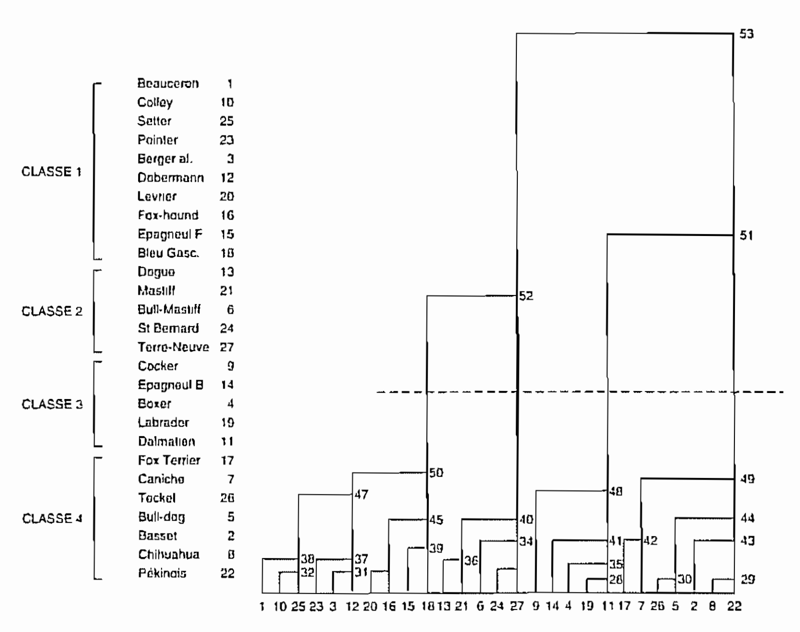


FIGURE 11.15

TABLEAU 11.2

CLASSIFICATION ASCENDANTE HIERARCHIQUE : DESCRIPTION DES NOEUDS

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
28	19	11	2	2.00	0.00000	*
29	8	22	2	2.00	0.00000	*
30	26	5	2	2.00	0.00000	*
31	10	1	2	2.00	0.01236	***
32	12	3	2	2.00	0.01236	***
33	16	20	2	2.00	0.01236	***
34	24	27	2	2.00	0.01236	***
35	28	4	3	3.00	0.01648	****
36	21	13	2	2.00	0.01759	****
37	15	25	2	2.00	0.01968	****
38	32	23	3	3.00	0.02060	****
39	18	33	3	3.00	0.03036	******
40	34	6	3	3.00	0.03119	******
41	35	14	4	4.00	0.03251	*******
42	7	17	2	2.00	0.03266	*******
43	29	2	3	3.00	0.03297	*******
44	37	31	4	4.00	0.04074	********
45	43	30	5	5.00	0.04698	*********
46	40	36	5	5.00	0.04939	**********
47	41	9	5	5.00	0.06935	*************
48	39	44	7	7.00	0.07612	***************
49	45	42	7	7.00	0.07898	***************
50	38	48	10	10.00	0.08497	****************
51	46	50	15	15.00	0.22780	***
52	49	47	12	12.00	0.27570	***
53	52	51	27	27.00	0.43314	***

SOMME DES INDICES DE NIVEAU = 1.66667

Représentation des individus et des centres de gravité des classes dans le premier plan factoriel

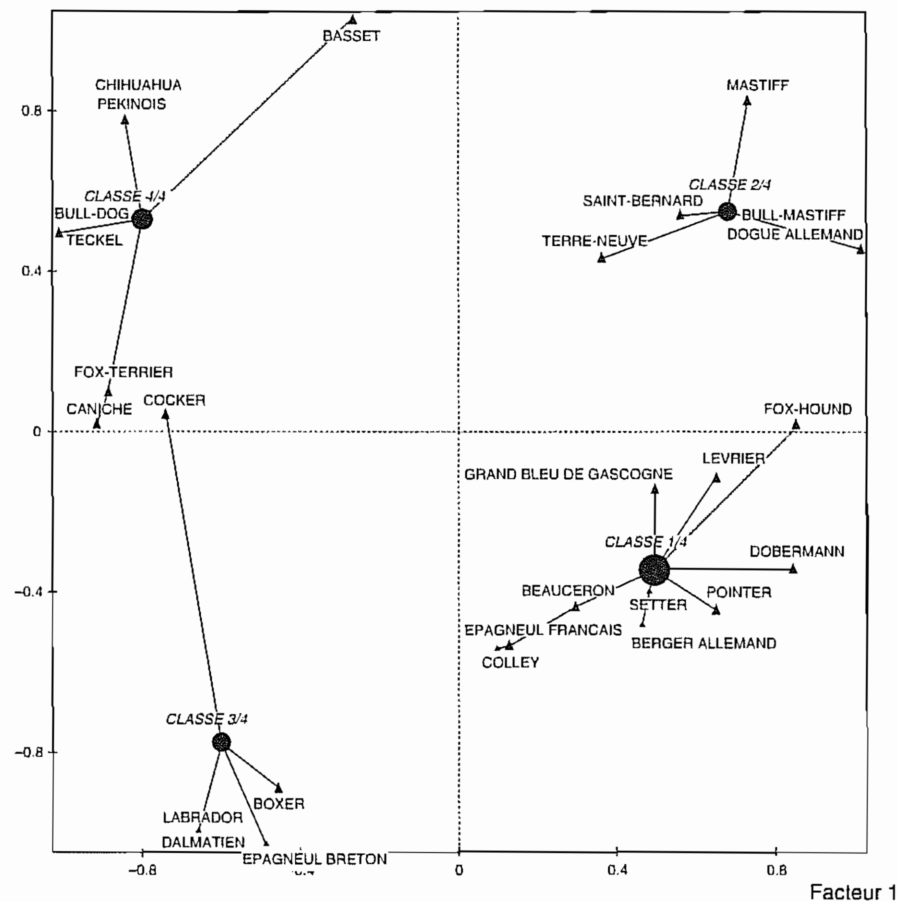


FIGURE 11.16

TROISIÈME PARTIE

Statistique inférentielle

12 Distributions des caractéristiques d'un échantillon

Le problème central de l'inférence statistique est rappelons-le, le suivant : disposant d'observations sur un échantillon de taille n on désire en déduire les propriétés de la population dont il est issu. Ainsi on cherchera à estimer, par exemple, la moyenne m de la population à partir de la moyenne $\bar{x}$ d'un échantillon. Ceci n'est possible que si l'échantillon a été tiré selon des règles rigoureuses destinées à en assurer la « représentativité » (voir chapitre 20). Le mode de tirage le plus simple et aussi le plus important est l'***échantillonnage aléatoire simple*** correspondant à des tirages équiprobables et indépendants les uns des autres. Dans ces conditions les observations deviennent des variables aléatoires ainsi que les résumés numériques usuels : il convient donc d'en chercher les lois de probabilité avant de tenter d'extrapoler à la population.

■ **Exemple :** On prélève au hasard n ampoules électriques dans une production et on mesure leurs durées de fonctionnement. Si les caractéristiques de fabrication n'ont pas varié d'une ampoule à l'autre, les différences entre les x_i peuvent être considérées comme des fluctuations de nature aléatoire. ■

Cette dernière remarque justifie l'hypothèse fondamentale de la théorie de l'échantillonnage : les valeurs observées x_i sont des réalisations d'une même variable aléatoire X, appelée variable parente. Dans notre exemple, ceci revient à postuler l'existence d'une variable abstraite, la durée de vie d'une ampoule de type donné, fabriquée dans des conditions données.

On peut cependant introduire aussi le modèle suivant : à chaque individu i tiré, on associe une variable aléatoire X_i dont on observe une seule réalisation x_i (exemple : X_i est la durée de vie de l'ampoule n° i qui, une fois l'expérience faite, a pris la valeur x_i).

L'hypothèse formulée plus haut revient alors à dire que les X_i sont des variables aléatoires ayant toutes la même distribution, celle de X. Pour des raisons de commodité, on supposera généralement les X_i mutuellement indépendantes (dans certains cas, l'indépendance deux à deux sera suffisante).

On a donc la double conception suivante, qui est à la base de la statistique mathématique : les valeurs observées $(x_1, x_2, \ldots, x_n)$ constituent n réalisations indépendantes d'une variable aléatoire X ou encore, une réalisation unique du n-uple $(X_1, X_2, \ldots, X_n)$ où les X_i sont n variables aléatoires indépendantes et de même loi.

Par extension, nous appellerons désormais échantillon le n-uple de variables aléatoires $(X_1, X_2, \ldots, X_n)$.

La théorie de l'échantillonnage se propose d'étudier les propriétés du n-uple $(X_1, X_2, \ldots, X_n)$ et des caractéristiques le résumant, encore appelées statistiques, à partir de la distribution supposée connue de la variable parente X, et d'étudier en particulier ce qui se passe lorsque la taille de l'échantillon est élevée.

Il est d'usage de résumer les n valeurs d'un échantillon $x_1, x_2, \ldots, x_n$ par quelques caractéristiques simples telles que moyenne, plus grande valeur, etc. Ces caractéristiques sont elles-mêmes des réalisations de variables aléatoires issues de $X_1, X_2, \ldots, X_n$.

DÉFINITION

Une statistique T est une variable aléatoire fonction mesurable de $X_1, X_2, \ldots, X_n$. $T = f(X_1, X_2, \ldots, X_n)$.

Une statistique peut être à valeurs dans $\mathbb{R}$ ou $\mathbb{R}^p$; dans le cas de $\mathbb{R}^p$, on parlera de statistique vectorielle.

Les premiers paragraphes de ce chapitre sont consacrés au cas des échantillons d'une variable aléatoire réelle. On donnera ensuite quelques résultats concernant les échantillons de vecteurs aléatoires.

12.1 FONCTION DE RÉPARTITION D'UN ÉCHANTILLON, STATISTIQUES D'ORDRE ET QUANTILES

12.1.1 Fonction de répartition empirique d'un échantillon

Désignons par $F_n^*(x)$ la proportion des n variables $X_1, X_2, \ldots, X_n$ qui sont inférieures à x.

$F_n^*(x)$ est donc une variable aléatoire pour tout x qui définit ainsi une fonction aléatoire appelée fonction de répartition empirique de l'échantillon, dont les réalisations sont des fonctions en escalier de sauts égaux à $1/n$ (fig. 12.1).

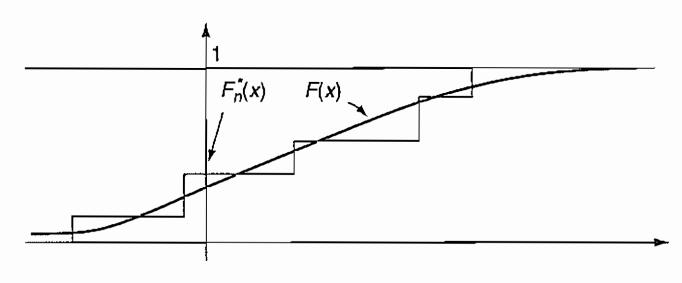


FIGURE 12.1

Si les x_i sont ordonnés par valeurs croissantes :

$$F_n^*(x) = 0 \qquad \text{si } x < x_1$$

$$F_n^*(x) = \frac{i-1}{n} \qquad \text{si } x_{i-1} \leq x < x_i$$

$$F_n^*(x) = 1 \qquad \text{si } x \geq x_n$$

12.1.2 Convergence de $F_n^*(x)$ vers $F(x)$

Ces trois théorèmes sont fondamentaux et justifient l'usage des échantillons en statistique.

Théorème 1

Pour tout x, on a $F_n^*(x) \xrightarrow{ps} F(x)$.

Démonstration : A x fixé, soit Y le nombre aléatoire de valeurs inférieures à x, qui est une somme de variables de Bernoulli de paramètre $F(x)$. D'après ce qui précède $F_n^*(x)$ qui n'est autre que Y/n converge presque sûrement vers la probabilité $F(x)$. ▪

Théorème 2 (Glivenko-Cantelli)

La convergence de F_n^* vers F est presque sûrement uniforme, c'est-à-dire que :

$$D_n = \sup_x |F_n^*(x) - F(x)| \rightarrow 0$$

Démonstration : voir Renyi, chapitre 7, p. 378. ▪

Théorème 3 (Kolmogorov)

$$\lim_{n \to \infty} P(\sqrt{n}D_n < y) = K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2)$$

Ce théorème signifie que la distribution asymptotique de la variable aléatoire D_n est connue et ne dépend pas de la variable de départ X, et permet de calculer des limites pour les valeurs de D_n. La loi exacte de la variable D_n a été tabulée (table A1.14 du recueil).

12.1.3 Échantillons ordonnés et lois des valeurs extrêmes

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon d'une variable aléatoire X. Les réalisations $x_1, x_2, \ldots, x_n$ peuvent être réordonnées en $y_1, y_2, \ldots, y_n$ où $y_1 < y_2 < \ldots < y_n$, les y_i constituent une permutation particulière des x_i. Les y_i sont des réalisations du n-uple de variables aléatoires $(Y_1, Y_2, \ldots, Y_n)$ qui constitue l'échantillon ordonné de X. Soit $F(x)$ la fonction de répartition de X de densité $f(x)$ et H_k et h_k les fonctions de répartition et densité de Y_k.

12.1.3.1 Loi de $Y_1 = \inf X_i$

On a $P(Y_1 < y) = 1 - P(Y_1 > y)$ et $P(\inf X_i > y) = \prod_{i=1}^{n} P(X_i > y)$ donc :

$$\boxed{\begin{aligned} H_1(y) &= 1 - [1 - F(y)]^n \\ h_1(y) &= n[1 - F(y)]^{n-1} f(y) \end{aligned}}$$

12.1.3.2 Loi de $Y_n = \sup X_i$

$$P(Y_n < y) = \prod_{i=1}^{n} P(X_i < y)$$

$$\boxed{\begin{aligned} H_n(y) &= [F(y)]^n \\ h_n(y) &= n[F(y)]^{n-1} f(y) \end{aligned}}$$

Ces deux lois servent en particulier pour la détection des valeurs « aberrantes » dans un échantillon : valeurs « trop » petites ou « trop » grandes.

Exemple : On sait que pour une loi LG(m ; σ) il y a une probabilité 1.35‰ de dépasser $m + 3\sigma$. Sur un échantillon de 100 observations la probabilité qu'il y en ait au moins une qui dépasse $m + 3\sigma$ monte à $1 - (0.99865)^{100} = 0.126$. Si inversement on cherche quelle est la valeur que Y_n a une probabilité 1.35‰ de dépasser on trouve : $F(y_n) = (0.99865)^{1/n}$ soit pour $n = 100$ environ $m + 4.3\sigma$.

12.1.3.3 Loi de l'étendue W

$$W = Y_n - Y_1$$

La loi du couple (Y_1, Y_n) s'obtient en écrivant :

$$\begin{aligned} P((Y_1 < y_1) \cap (Y_n < y_n)) &= P(Y_n < y_n) - P((Y_n < y_n) \cap (Y_1 > y_1)) \\ &= (F(y_n))^n - (F(y_n) - F(y_1))^n \end{aligned}$$

d'où la densité du couple $Y_1 Y_n$ en dérivant deux fois :

$$h(y_1, y_n) = n(n-1)(F(y_n) - F(y_1))^{n-2} f(y_n) f(y_1)$$

Avec le changement de variables $(Y_1, Y_n) \rightarrow (Y_1, W)$ on obtient la fonction de répartition de W :

$$G(w) = \int_{\mathbb{R}} n[F(x + w) - F(x)]^{n-1} f(x)\mathrm{d}x$$

et sa densité : $$g(w) = n(n-1)\int_{\mathbb{R}} [F(x + w) - F(x)]^{n-2} f(x) f(x + w)\mathrm{d}x$$

On trouve alors : $$E(W) = \int_{\mathbb{R}} (1 - (F(x))^n - (1 - F(x))^n)\mathrm{d}x$$

en intégrant par parties $E(Y_n) - E(Y_1)$.

12.1.3.4 Loi de Y_k

Appelons $R_n(x)$ le nombre de répétitions de l'événement $X < x$ en n expériences indépendantes, qui suit donc une loi binomiale :

$$P(R_n(x) = i) = C_n^i[F(x)]^i[1 - F(x)]^{n-i}$$

L'événement $Y_k < x$ peut être obtenu de plusieurs manières, soit que les k premières valeurs de X soient inférieures à x et elles seules, soit qu'il y en ait $k + 1$, etc.

Donc :

$$P(Y_k < x) = \sum_{i=k}^{n} C_n^i[F(x)]^i[1 - F(x)]^{n-i}$$

L'événement $x < Y_k < x + dx$ se réalise si un des x_i est compris entre x et $x + dx$, si $(k - 1)x_i$ sont inférieurs à x et si les $n - k$ restant sont supérieurs à x. Les probabilités respectives de ces différents événements sont $f(x)dx$, $[F(x)]^{k-1}$, $[1 - F(x)]^{n-k}$. Il y a n manières de réaliser le premier événement et C_{n-1}^{k-1} manières de réaliser les deux autres (C_{n-1}^{k-1} façons de choisir les x_i inférieurs à x, les autres étant alors supérieurs) :

$$\boxed{h_k(x) = nC_{n-1}^{k-1}[F(x)]^{k-1}[1 - F(x)]^{n-k}f(x)}$$

$F(Y_k)$ suit donc une loi bêta I de paramètres k et $n - k + 1$.

12.1.3.5 Résultats asymptotiques pour les extrêmes

L'étude du comportement de Y_1 et Y_n lorsque $n \to \infty$ est l'objet de la théorie des valeurs extrêmes dont nous donnons ci-dessous quelques résultats. Nous nous bornerons à étudier Y_n puisque $Y_1 = -\sup(-X_1, -X_2, \ldots, -X_n)$.

Remarquons que si $n \to \infty$:

$$(F(y))^n \to 0 \qquad \text{si } F(y) < 1$$
$$(F(y))^n \to 1 \qquad \text{si } F(y) = 1$$

ce qui est sans intérêt. Il convient plutôt de rechercher s'il existe des coefficients a_n et b_n tels que $a_nY_n + b_n$ tende vers une limite non dégénérée, par une opération semblable au centrage-réduction dans le théorème central-limite.

La méthode est la suivante : soit $G(y)$ la loi limite de $a_nY_n + b_n$. Puisque la plus grande des Nn valeurs $X_1, X_2, \ldots, X_{Nn}$ est aussi la plus grande des N maxima suivants : $\sup(X_1, X_2, \ldots, X_n)$; $\sup(X_{n+1}, \ldots, X_{2n})$; ... ; $\sup(X_{(N-1)n}$; ... $X_{Nn})$ on doit avoir :

$$(G(y))^N = G(a_Ny + b_N)$$

On démontre alors que les seules solutions de cette équation fonctionnelle sont les suivantes pour X non borné :

- type I : $G(y) = \exp(-\exp(-y))$ loi de Gumbel obtenue si $1 - F(x)$ tend vers 0 comme $\exp(-x)$ quand $x \to \infty$;
- type II : $G(y) = \exp(-y^{\alpha})$; $y > 0$ loi de Weibull (ou de Fréchet) si $1 - F(x)$ tend vers 0 comme x^{-k} quand $x \to \infty$ (voir chapitre 2, paragr. 2.3.8 et 2.3.9).

Ceci permet en pratique de pouvoir faire les approximations suivantes si n est très grand :

$$H_n(y) \simeq \exp\left(-\exp\left(-\left(\frac{y-\xi}{\theta}\right)\right)\right) \quad \text{ou} \quad \exp\left(-\left(\frac{x-a}{b}\right)^{\alpha}\right)$$

12.1.3.6 Distributions asymptotiques des quantiles

Si F est continue, rappelons que le quantile d'ordre p noté q_p est la valeur de x telle que $F(x) = p$. Le quantile empirique d'un n-échantillon Q_p est égal à $Y_{[np]+1}$ où $[np]$ est la partie entière de np supposé non entier.

On démontre (voir Fourgeaud-Fuchs, 1972) que si $n \to \infty$:

$$\sqrt{n}(Q_p - q_p) \to \text{LG}\left(q_p\,;\frac{\sqrt{p(1-p)}}{f(q_p)}\right)$$

D'où en particulier pour la médiane :

$$\sqrt{n}(Q_{1/2} - q_{1/2}) \to \text{LG}\left(q_{1/2}\,;\frac{1}{2f(q_{1/2})}\right)$$

12.2 DISTRIBUTIONS D'ÉCHANTILLONNAGE DE CERTAINS MOMENTS

12.2.1 Étude de la statistique $\overline{X}$

DÉFINITION

La statistique $\overline{X}$ ou moyenne empirique de l'échantillon est :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

12.2.1.1 Propriétés élémentaires

Soit m et σ l'espérance et l'écart-type de la variable parente ; on a alors :

$$E(\overline{X}) = m \quad \text{et} \quad V(\overline{X}) = \frac{\sigma^2}{n}$$

Démonstration :

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}nm = m$$

$$V(\overline{X}) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

d'après l'indépendance des X_i.

Si μ_3 et μ_4 sont les moments centrés d'ordre 3 et 4 de X on a :

$$\mu_3(\overline{X}) = \frac{\mu_3}{n^2} \quad \text{et} \quad \mu_4(\overline{X}) = \frac{\mu_4 + 3\sigma^4(n-1)}{n^3}$$

On en déduit :

$$\boxed{\gamma_1(\overline{X}) = \frac{\gamma_1}{\sqrt{n}} \quad \text{et} \quad \gamma_2(\overline{X}) = 3 + \frac{\gamma_2 - 3}{n}}$$

où γ_1 et γ_2 sont les coefficients d'asymétrie et d'aplatissement de X.

Lorsque $n \to \infty$, $V(\overline{X}) \to 0$, il s'ensuit que $\overline{X}$ converge en moyenne quadratique vers m puisque $E[(\overline{X} - m)^2] \to 0$.

Ce dernier résultat est une forme des lois des grands nombres que nous allons énoncer sous un aspect plus général.

On voit de plus que si $n \to \infty$, $\gamma_1(\overline{X}) \to 0$ et $\gamma_2(\overline{X}) \to 3$, ce qui traduit la normalité asymptotique de $\overline{X}$. ■

12.2.1.2 Lois des grands nombres

Elles sont de deux types : lois faibles mettant en jeu la convergence en probabilité et lois fortes relatives à la convergence presque sûre.

Nous considérons ici des suites de variables aléatoires $X_1, X_2, \ldots, X_n$ non nécessairement de même loi.

- Loi faible des grands nombres

Soit $X_1, X_2, \ldots, X_n$ indépendantes d'espérance $m_1, m_2, \ldots, m_n$ finies et de variance $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ finies.

Si $\frac{1}{n}\sum_{i=1}^{n} m_i \to m$ et si $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \to 0$, alors $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est tel que :

$$\overline{X} \xrightarrow{\mathcal{P}} m$$

- Loi forte des grands nombres

Soit $X_1, X_2, \ldots, X_n$ indépendantes telles que $\frac{1}{n}\sum_{i=1}^{n} m_i \to m$ et $\sum_{i=1}^{\infty}\frac{\sigma_i^2}{i^2}$ est convergente ; alors :

$$\overline{X} \xrightarrow{\text{ps}} m$$

(Pour la démonstration, *cf.* Renyi, chapitre 7).

Application : Cas des échantillons : on voit aisément que $\overline{X} \xrightarrow{\text{ps}} m$ car la condition $\sum_1^{\infty} \frac{\sigma_i^2}{i^2}$ convergente est réalisée puisque :

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} = \sigma^2 \sum_{i=1}^{\infty} \frac{1}{i^2}$$

et l'on sait que la série $\sum \frac{1}{i^2}$ converge.

- Distribution

Le théorème central-limite établi au chapitre 2 peut s'écrire :

$$\boxed{\frac{\overline{X} - m}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} U\ \mathrm{LG}(0,\ 1)}$$

Il suffit en effet de poser : $X_1 + X_2 + \cdots + X_n = n\overline{X}$.

Ce résultat est d'une importance capitale en statistique.

12.2.1.3 Application : loi d'un pourcentage

On prélève indépendamment et avec remise n individus d'une population séparée en deux sous-populations A et $\overline{A}$ de proportions p et $1 - p$ (pièces défectueuses ou correctes dans une production industrielle par exemple).

Soit K le nombre d'individus de la sous-population A obtenus dans l'échantillon. On sait que K suit une loi binomiale $\mathcal{B}(n\ ;\ p)$.

Notons $F = K/n$ la fréquence empirique de la catégorie A.

F est la moyenne arithmétique de n variables de Bernoulli de paramètre p indépendantes.

On a donc :

$$E(F) = p$$
$$V(F) = \frac{p(1 - p)}{n}$$

et si n est grand $F \simeq \mathrm{LG}\left(p : \sqrt{\frac{p(1-p)}{n}}\right)$ en raison du théorème central-limite.

La convergence de F vers p, connue sous le nom de théorème de De Moivre-Laplace, est une des premières applications de la loi des grands nombres. Ce résultat a inspiré la théorie fréquentiste des probabilités (voir chapitre 1).

Application numérique : Comme pour la loi binomiale l'approximation gaussienne de F est valable si np et $n(1 - p)$ sont tous deux supérieurs à 5.

Ainsi pour un échantillon de 400 pièces issues d'une fabrication où 10 % sont défectueuses, on peut s'attendre à trouver dans 95 % des cas un pourcentage de défectueux dans l'échantillon compris entre 10 % $\pm 1.96\sqrt{\frac{0.10 \times 0.90}{400}}$, soit $9.7\ \% < F < 10.3\%$.

12.2.2 Étude de la statistique S^2

DÉFINITION

La statistique S^2 ou variance empirique d'échantillon est :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

12.2.2.1 Propriétés

$$S^2 = \frac{1}{n}\left(\sum_{i=1}^{n}X_i^2\right) - (\overline{X})^2$$

Démonstration : Il suffit de développer.

- *Convergence presque sûre de S^2 vers σ^2*

D'après les lois des grands nombres :

$$\frac{1}{n}\left(\sum_{i=1}^{n}X_i^2\right) \xrightarrow{\text{ps}} E(X^2)$$

et :

$$\overline{X}^2 \longrightarrow [E(X)]^2$$

donc :

$$S^2 \xrightarrow{\text{ps}} E(X^2) - [E(X)]^2 = \sigma^2$$

- *Décomposition de S^2*

Partons de $X_i - m = X_i - \overline{X} + \overline{X} - m$.

On a alors : $\sum_{i=1}^{n}(X_i - m)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(\overline{X} - m)^2 + 2(\overline{X} - m)\sum_{i=1}^{n}(X_i - \overline{X})$

Comme $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$, on trouve :

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 + (\overline{X} - m)^2$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X} - m)^2$$

- *Biais de S^2*

THÉORÈME

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

Ce théorème montre que $E(S^2) \neq \sigma^2$. On dit que S^2 est une statistique biaisée pour σ^2.

Démonstration :

$$E(S^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i - m)^2 - E(\overline{X} - m)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} V(X_i) - V(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n}$$

Le biais vaut σ^2/n et tend donc vers 0.

• *Variance de S^2*

Un calcul dont la longueur est la seule difficulté montre que :

$$V(S^2) = \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4]$$

et si $n \to \infty$:

$$V(S^2) \sim \frac{\mu_4 - \sigma^4}{n}$$

La variance S^2 étant biaisée et ayant donc tendance à sous-estimer σ^2, on utilise fréquemment la ***variance corrigée*** dont l'espérance vaut exactement σ^2 :

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$E(S^{*2}) = \sigma^2$$

Cependant, l'écart-type corrigé S^* reste biaisé pour σ car :

$$E(\sqrt{S^{*2}}) \neq \sqrt{E(S^{*2})}$$

mais est asymptotiquement sans biais.

Il n'existe pas d'expression générale donnant $E(S^*)$ pour toute distribution. On verra plus loin une formule exacte dans le cas où les X_i suivent des lois normales.

12.2.2.2 Théorème limite pour S^2

$$\frac{S^2 - \frac{n-1}{n}\sigma^2}{\sqrt{V(S^2)}} \xrightarrow{\mathcal{L}} U \in \text{LG}(0, 1)$$

ce qui peut s'écrire avec l'approximation précédente :

$$\frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}}\sqrt{n} \xrightarrow{\mathcal{L}} U \in \text{LG}(0, 1)$$

12.2.2.3 Corrélation entre $\overline{X}$ et S^2

Cherchons $\text{cov}(\overline{X}, S^2)$:

$$\text{cov}(\overline{X}, S^2) = E\left[(\overline{X} - m)\left(S^2 - \frac{n-1}{n}\sigma^2\right)\right]$$

Nous pouvons supposer sans nuire à la généralité que $m = 0$, car on sait que la covariance est insensible à un changement par translation sur un des termes :

$$\text{cov}(\overline{X}, S^2) = E(\overline{X}S^2)$$

$$E(\overline{X}S^2) = E\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}X_j^2 - \overline{X}^2\right)\right]$$

$$= \frac{1}{n^2}E[(\sum X_i)(\sum X_j^2)] - E(\overline{X}^3)$$

$$= \frac{1}{n^2}E\left[\sum_i\sum_j X_iX_j^2\right] - E(\overline{X}^3)$$

$$= \frac{1}{n^2}E\left(\sum_i X_i^3\right) - \frac{1}{n^3}E\left(\sum_i X_i^3\right)$$

car $E(X_iX_j^2) = 0$ pour $i \neq j$ à cause de l'indépendance :

$$E(\overline{X}, S^2) = \frac{\mu_3}{n} - \frac{\mu_3}{n^2} = \frac{n-1}{n^2}\mu_3$$

Avec la formule établie précédemment pour $V(S^2)$, on trouve que :

$$\rho(\overline{X}\,;S^2) = \frac{\mu_3}{\sigma\sqrt{\mu_4 - \dfrac{n-3}{n-1}\sigma^4}}$$

et n'est donc nul que si μ_3 est nul, ce qui est le cas des distributions symétriques.

Il faut se garder de passer de la non corrélation à l'indépendance et nous verrons dans un paragraphe suivant que $\overline{X}$ et S^2 ne sont indépendants que si X suit une loi de Laplace-Gauss.

12.2.3 Cas des échantillons gaussiens

On suppose maintenant que $X \in \text{LG}(m, \sigma)$:

12.2.3.1 Loi de $\overline{X}$

$\overline{X}$ combinaison linéaire de variables de Laplace-Gauss est aussi de Laplace-Gauss et

$$\overline{X} \in \text{LG}\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

Il s'agit ici d'une loi exacte.

12.2.3.2 Loi de S^2 et indépendance entre $\overline{X}$ et S^2

D'après la décomposition de S^2 on peut écrire :

$$\sum_{i=1}^{n}(X_i - m)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - m)^2$$

Divisons par σ^2 de chaque côté :

$$\sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = \frac{nS^2}{\sigma^2} + \left(\frac{\overline{X} - m}{\sigma/\sqrt{n}}\right)^2$$

Nous sommes dans les conditions d'application du théorème de Cochran.

Le premier membre est une somme de n carrés de variables centrées réduites et suit donc un χ^2_n. Le deuxième membre est constitué de la somme de deux formes quadratiques sur ces variables de rang 1 pour $\left(\frac{\overline{X} - m}{\sigma/\sqrt{n}}\right)^2$ de rang $n - 1$ pour $\frac{nS^2}{\sigma^2}$: en effet $\overline{X}$ est lié aux X_i et l'on a la relation $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$.

On en déduit les deux résultats suivants :

THÉORÈME 1

$$\frac{nS^2}{\sigma^2} \textit{ suit une loi de } \chi^2_{n-1}$$

THÉORÈME 2

$$\overline{X} \textit{ et } S^2 \textit{ sont indépendants}$$

On peut de plus démontrer la réciproque du théorème 2 : si $\overline{X}$ et S^2 sont indépendants alors X est LG (m, σ), il s'agit donc d'une propriété caractéristique.

Application : Puisque $\frac{\overline{X} - m}{\sigma}\sqrt{n} \in \text{LG}(0, 1)$ et $\frac{nS^2}{\sigma^2} \in \chi^2_{n-1}$ on aura :

$$T_{n-1} = \frac{\frac{\overline{X} - m}{\sigma}\sqrt{n}}{\sqrt{\frac{nS^2}{(n-1)\sigma^2}}} = \frac{\overline{X} - m}{S}\sqrt{n-1}$$

où T_{n-1} est une variable de Student à $n - 1$ degrés de liberté.

Ce résultat est extrêmement utile car il ne dépend pas de σ et servira donc chaque fois que σ est inconnu.

Exemple : On prélève 25 pièces dans une production industrielle. Une étude préalable a montré que le diamètre de ces pièces suivait une loi gaussienne LG (10 ; 2). Entre quelles valeurs a-t-on 90 chances sur 100 de trouver le diamètre moyen de ces 25 pièces et leur écart-type ?

$$\overline{X}\ \text{LG}\left(10 ; \frac{2}{\sqrt{25}}\right)$$

avec une probabilité 0.90 on trouvera $10 - 1.64\frac{2}{\sqrt{25}} < \overline{X} < 10 + 1.64\frac{2}{\sqrt{25}}$ soit $9.34 < \overline{X} < 10.66$ car pour la variable centrée-réduite U : $P(-1.64 < U < 1.64) = 0.9$.

Comme $\frac{nS^2}{\sigma^2} = \chi^2_{n-1}$ on se reporte aux tables de la loi du χ^2_{24}. En prenant conventionnellement des risques d'erreur symétriques on trouve :

$$13.848 < \frac{25S^2}{4} < 36.415 \quad \text{(fig.12.2)}$$

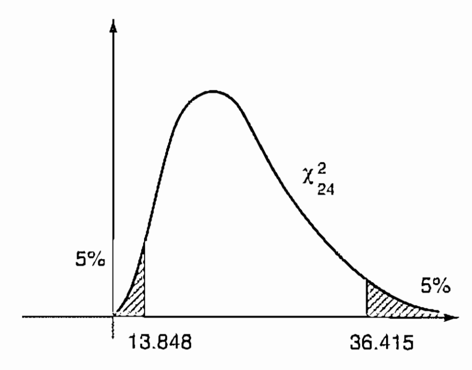


FIGURE 12.2

soit :

$$\frac{2}{5}\sqrt{13.848} < S < \frac{2}{5}\sqrt{36.415}$$

d'où $1.49 < S < 2.41$.

12.2.3.3 Espérance et variance des principales caractéristiques d'un échantillon gaussien

Le tableau 12.1 récapitule les résultats :

$$\gamma_1 = \frac{1/n\sum_{i=1}^{p}(x_i - \bar{x})^3}{s^3} \qquad \gamma_2 = \frac{1/n\sum_{i=1}^{p}(x_i - \bar{x})^4}{s^4}$$

TABLEAU 12.1

Statistique	Espérance	Variance
$\overline{X}$	m	σ^2/n
S^2	$\frac{n-1}{n}\sigma^2$	$\frac{2(n-1)}{n^2}\sigma^4$
S^{*2}	σ^2	$\frac{2\sigma^4}{n-1}$
S^*	$c_4\,\sigma$	$(1-c_4^2)\sigma^2$
R	$d_2\,\sigma$	$d_3^2\,\sigma^2$
γ_1	$\simeq 0$	$\simeq \frac{6}{n}$
γ_2	$\simeq 3$	$\simeq \frac{24}{n}$
Médiane	m	$\simeq \frac{\pi}{2}\frac{\sigma^2}{n}$

Un calcul d'intégrale permet de trouver la valeur de $E(S^*)$:

$$E(S^*) = \sigma\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

que l'on exprime souvent sous la forme $E(S^*) = c_4\sigma$, où c_4 tend vers 1 quand n augmente (table A18). S^* est donc asymptotiquement sans biais pour σ.

On en déduit aisément la variance $V(S^*) = E(S^{*2}) - (E(S^*))^2 = \sigma^2 - (c_4\sigma)^2 = \sigma^2(1 - c_4^2)$

Pour l'étendue R, les calculs ne sont pas aisés : les coefficients usuellement notés d_2 et d_3 qui permettent de calculer $E(R) = d_2\sigma$ et $V(R) = (d_3\sigma)^2$ figurent dans la table A18.

On notera que quand n augmente, d_2 tend vers l'infini, car la loi normale a pour support l'ensemble des nombres réels.

12.2.4 Application aux cartes de contrôle

Il s'agit d'une des plus importantes applications industrielles directes de la théorie de l'échantillonnage.

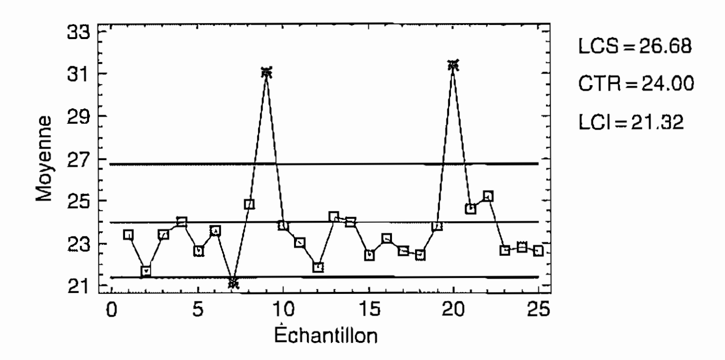
Introduites par W.A. Shewhart dès 1931, les cartes de contrôle permettent de suivre au cours du temps la moyenne et la dispersion d'un procédé de fabriacation afin de détecter des écarts significatifs (déréglages ou dérives) par rapport aux valeurs nominales ou consignes à respecter. En effet, tout procédé est soumis à des variations, que l'on modélise souvent par une loi normale : par exemple le diamètre de pièces mécaniques suit une loi N(m, σ). Soit m_0 et σ_0 les valeurs nominales.

On prélève à intervalles réguliers des échantillons de n pièces. La carte de Shewhart ($\bar{X}$; S) est un double graphique où l'on reporte les valeurs successives de la moyenne et de l'écart-type corrigé de chaque échantillon. La ligne centrale correspond à l'espérance de la statistique si le procédé est bien réglé : $E(\bar{X}) = m_0$ $E(S^*) = c_4\sigma_0$. Les limites de contrôle sont conventionnellement à ± 3 écart-types de la valeur centrale soit :

$$m_0 \pm 3\frac{\sigma_0}{\sqrt{n}} \quad \text{et} \quad c_4\sigma_0 \pm 3\sqrt{(1 - c_4^2)}\sigma_0$$

Pour $n \leq 5$ la limite inférieure de contrôle pour S est mise à zéro, pour éviter une valeur négative. La probabilité de sortir des limites de contrôle étant très faible lorsque le procédé est bien réglé, on interviendra dès que l'une des deux statistiques sort des limites.

La figure suivante illustre une carte de contrôle pour un procédé où $m_0 = 24$ et $\sigma_0 = 2$ avec des échantillons de taille 5. Des interventions auraient du avoir lieu aux instants 7, 9 et 20, car la moyenne était sortie des limites.



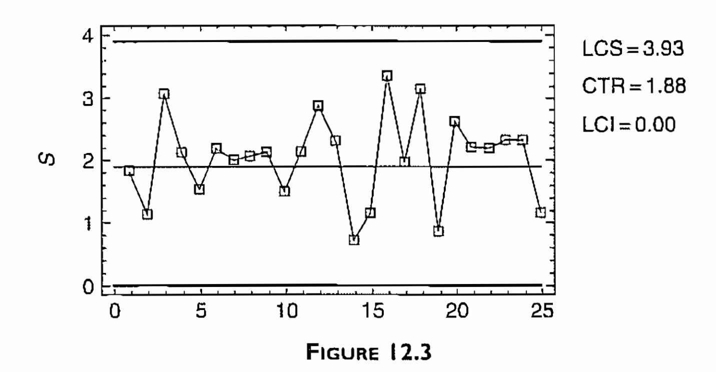


FIGURE 12.3

Il existe bien d'autres cartes de contrôle, *cf.* L. Jaupi, 2002.

12.3 DISTRIBUTION DU CENTRE DE GRAVITÉ ET DE LA MATRICE DE VARIANCE D'UN ÉCHANTILLON GAUSSIEN p-DIMENSIONNEL

Soit un échantillon de taille n de la loi $N_p(\boldsymbol{\mu} ; \boldsymbol{\Sigma})$ (c'est-à-dire un tableau de données à n lignes et p colonnes), il suffit alors d'appliquer les résultats du chapitre 4 pour obtenir que :

$$\sqrt{n}\mathbf{g} \sim N_p(\sqrt{n}\boldsymbol{\mu} ; \boldsymbol{\Sigma}) \qquad \text{soit} \quad \mathbf{g} \sim N_p\left(\boldsymbol{\mu} ; \frac{1}{n}\boldsymbol{\Sigma}\right)$$

La matrice de variance $\mathbf{V}$ suit alors une loi de Wishart à $(n - 1)$ degrés de liberté :

$$n\mathbf{V} \sim W_p(n - 1 ; \boldsymbol{\Sigma})$$

$\mathbf{V}$ et $\mathbf{g}$ sont des statistiques indépendantes.

La distance de g à μ au sens de Mahalanobis est :

$$(\mathbf{g} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{g} - \boldsymbol{\mu})$$

et on a :

$$n(\mathbf{g} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{g} - \boldsymbol{\mu}) \sim \chi^2_p$$

Le résultat suivant est cependant plus utile car il ne fait intervenir que la matrice **V** observée et non la matrice $\boldsymbol{\Sigma}$ théorique :

$$(n - 1)(\mathbf{g} - \boldsymbol{\mu})' \mathbf{V}^{-1}(\mathbf{g} - \boldsymbol{\mu}) = T^2_p(n - 1)$$

soit :

$$\boxed{\frac{n-p}{p}(\mathbf{g} - \boldsymbol{\mu})' \mathbf{V}^{-1}(\mathbf{g} - \boldsymbol{\mu}) = F(p\,;n - p)}$$

il étend la formule unidimensionnelle du T de Student.

12.4 LA MÉTHODE « DELTA » ET LES STATISTIQUES ASYMPTOTIQUEMENT NORMALES

Soit T une statistique telle que si $n \to \infty$ $T \to \text{LG}\left(\theta\,;\frac{\sigma(\theta)}{\sqrt{n}}\right)$ et g une fonction dérivable. Alors $g(T)$ est également une statistique asymptotiquement normale et $T \to \text{LG}\left(g(\theta)\,;\frac{g'(\theta)\sigma(\theta)}{\sqrt{n}}\right)$.

En effet d'après la formule des accroissements finis : $g(T) - g(\theta) = (T - \theta)g'(\theta) + \varepsilon$ où ε est ici une variable aléatoire qui tend vers 0 lorsque T tend vers θ, donc quand $n \to \infty$.

La distribution asymptotique de $g(T) - g(\theta)$ est donc celle de $g'(\theta)(T - \theta)$ et on a $V(g(T))$ # $(g'(\theta))^2\,V(T)$ d'où le résultat annoncé.

Ce résultat est particulièrement utile lorsque l'on veut obtenir une variance asymptotique indépendante de θ : il suffit de résoudre l'équation différentielle $g'(\theta)\sigma(\theta) = c$. En voici trois applications :

12.4.1 Stabilisation de la variance d'un pourcentage

On a vu que $F \to \text{LG}\left(p\,;\sqrt{\frac{p(1-p)}{n}}\right)$, d'où :

$$g(F) \to \text{LG}\left(g(p)\,;\frac{\sqrt{p(1-p)}\,g'(p)}{\sqrt{n}}\right)$$

Si $g'(p) = \dfrac{c}{\sqrt{p(1-p)}}$, il vient $g(p) = 2c \text{ Arc} \sin \sqrt{p} + K$. En prenant $c = 1/2$ et $K = 0$ on en déduit que :

$$\text{Arc} \sin \sqrt{F} \rightarrow LG\left(\text{Arc} \sin \sqrt{p} \,;\, \frac{1}{2\sqrt{n}}\right)$$

12.4.2 Stabilisation de la variance d'une loi de Poisson

Soit $X\, \mathscr{P}(\lambda)$. On sait que $X \rightarrow LG(\lambda \,;\, \sqrt{\lambda})$ d'où :

$$\sqrt{X} \rightarrow LG\left(\sqrt{\lambda} \,;\, \frac{1}{2}\right)$$

12.4.3 Valeurs propres d'une matrice de variance

Soit un n-échantillon d'une loi normale p-dimensionnelle $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $\mathbf{V}^* = \dfrac{n}{n-1}\mathbf{V}$ la matrice de variance corrigée de l'échantillon.

Si λ_i et l_i désignent les $i^{\text{ème}}$ valeurs propres de $\boldsymbol{\Sigma}$ et de $\mathbf{V}^*$ respectivement, T. W. Anderson a montré que $\sqrt{n-1}(l_i - \lambda_i)$ converge vers une loi normale $LG(0 \,;\, \lambda_i\sqrt{2})$.

On en déduit que $\ln l_i$ a pour distribution approchée une $LG\left(\ln \lambda_i \,;\, \sqrt{\dfrac{2}{n-1}}\right)$, ce qui permet d'écrire :

$$0.95 = P\left(\ln \lambda_i - 1.96\sqrt{\frac{2}{n-1}} < \ln l_i < \ln \lambda_i + 1.96\sqrt{\frac{2}{n-1}}\right)$$

d'où : $$l_i \exp\left(-1.96\sqrt{\frac{2}{n-1}}\right) < \lambda_i < l_i \exp\left(1.96\sqrt{\frac{2}{n-1}}\right) \quad (cf \text{ ch 7, § 7.3.2})$$

12.4.4 Généralisation au cas multidimensionnel

Si $\mathbf{X} \rightarrow N_p\left(\boldsymbol{\mu} \,;\, \dfrac{\boldsymbol{\Sigma}}{n}\right)$ et si $\mathbf{y} = \varphi(\mathbf{X})$ avec φ application de $\mathbb{R}^p$ dans $\mathbb{R}^q$ différentiable alors :

$$\varphi(\mathbf{X}) \rightarrow N_q\left(\varphi(\boldsymbol{\mu}) \,;\, \frac{\Delta \Sigma \Delta'}{n}\right)$$

où Δ est la matrice des dérivées partielles de φ au point $\boldsymbol{\mu}$.

Ce résultat est souvent utilisé pour calculer des intervalles de confiance asymptotiques pour des paramètres multidimensionnels, le nom de méthode delta provient de l'usage des dérivées.

13 L'estimation

13.1 GÉNÉRALITÉS

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (m ; σ, etc.) à l'aide d'un échantillon de n observations issues de cette population. On supposera vérifiée l'hypothèse d'échantillonnage aléatoire simple.

13.1.1 Exemples élémentaires

Les lois des grands nombres justifient l'usage de $\bar{x}$ et de s^2 comme estimations de m et σ^2 respectivement : on sait que $\overline{X} \xrightarrow{\text{ps}} m$ et $S^2 \xrightarrow{\text{ps}} \sigma^2$. De même, la fréquence empirique f d'un événement est une estimation de sa probabilité p.

Les variables aléatoires $\overline{X}$, S^2, F sont appelées alors ***estimateurs*** de m, σ^2, p respectivement.

Cependant le même paramètre peut être estimé à l'aide d'estimateurs différents : pour une distribution symétrique la médiane de l'échantillon est également une estimation de m.

Afin de choisir entre plusieurs estimateurs possibles d'un même paramètre il faut définir les qualités exigées d'un estimateur.

13.1.2 Qualités d'un estimateur

Soit θ le paramètre à estimer et T un estimateur, c'est-à-dire une fonction des X_i à valeurs dans un domaine acceptable pour θ.

La première qualité d'un estimateur est d'être ***convergent***. Il est souhaitable que si $n \to \infty$ $T \to \theta$. C'est le cas des estimateurs présentés au paragraphe précédent. Deux estimateurs convergents ne convergent cependant pas nécessairement à la même vitesse, ceci est lié, pour une taille d'échantillon donnée, à la notion de ***précision*** d'un estimateur.

Un estimateur est une variable aléatoire. Supposons connue sa loi de probabilité pour une valeur donnée de θ. La figure 13.1 illustre alors les deux composantes de l'erreur d'estimation.

L'erreur d'estimation $T - \theta$ qui est une variable aléatoire se décompose de façon élémentaire en $T - E(T) + E(T) - \theta$ ou $E(T)$ est l'espérance de l'estimateur.

$T - E(T)$ représente les fluctuations aléatoires de T autour de sa valeur moyenne tandis que $E(T) - \theta$ est assimilable à une erreur systématique due au fait que T varie autour de sa valeur centrale $E(T)$ et non autour de θ.

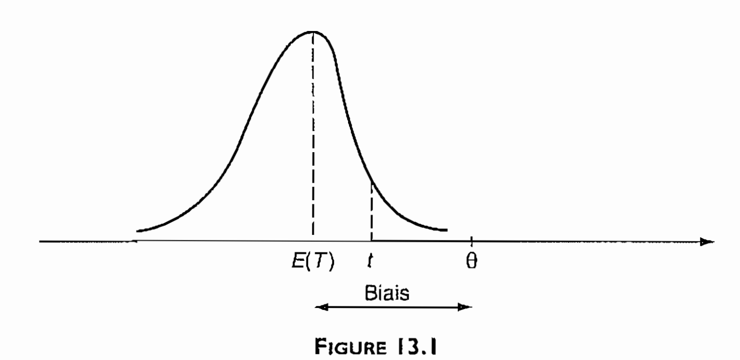


FIGURE 13.1

La quantité $E(T) - \theta$ s'appelle le ***biais***. Il est donc souhaitable d'utiliser des estimateurs ***sans biais***, tels que $E(T) = \theta$. Ainsi $\overline{X}$ est sans biais pour m, mais S^2 est biaisé pour σ^2.

Il est donc souvent préférable d'utiliser $S^{*2} = \dfrac{n}{n-1} S^2$ pour estimer σ^2.

On sait cependant que S^* n'est pas un estimateur sans biais de σ.

On mesure généralement la précision d'un estimateur T par l'***erreur quadratique moyenne*** :

$$E((T-\theta)^2)$$

On peut écrire :

$$E[(T-\theta)^2] = E[(T - E(T) + E(T) - \theta)^2] = E[(T-E(T))^2] + 2E[(T-E(T))(E(T)-\theta)] + E[(E(T)-\theta)^2]$$

Comme $E(T) - \theta$ est une constante et que $E[T - E(T)] = 0$ il vient :

$$\boxed{E[(T-\theta)^2] = V(T) + [E(T) - \theta]^2}$$

De deux estimateurs sans biais, le plus précis est donc celui de variance minimale.

Montrons ainsi que si m est connu l'estimateur $T = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ est meilleur que S^{*2} :

En effet : $$V(T) = \frac{1}{n^2} V\left(\sum_{i=1}^{n}(X_i - m)^2\right) = \frac{1}{n} V[(X-m)^2]$$

$$V(T) = \frac{1}{n}[E(X-m)^4 - [E(X-m)^2]^2] = \frac{1}{n}[\mu_4 - \sigma^4]$$

et : $$V(S^{*2}) = \left(\frac{n}{n-1}\right)^2 V(S^2) = \left(\frac{n}{n-1}\right)^2 \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4]$$

$$V(S^{*2}) = \frac{1}{n}\left[\mu_4 - \frac{n-3}{n-1}\sigma^4\right]$$

donc $V(T) < V(S^{*2})$.

13.1.3 Recherche du meilleur estimateur d'un paramètre θ

On ne peut résoudre d'une façon générale le problème de la recherche du meilleur estimateur d'un paramètre sans faire d'hypothèses sur le phénomène échantillonné. En effet la variance d'un estimateur ne peut en général se calculer que si l'on connaît la loi de T qui dépend de celle des X_i.

Le modèle utilisé en théorie classique de l'estimation est alors le suivant : on observe un échantillon d'une variable X dont on connaît la loi de probabilité à l'exception de la valeur numérique d'un ou de plusieurs paramètres (par exemple : X suit une loi de Poisson $\mathcal{P}(\theta)$ de paramètre θ inconnu). En d'autres termes la variable X est définie par une famille paramétrée de lois $f(x ; \theta)$ où f a une expression analytique connue.

Cependant la théorie de l'estimation ne permet pas de résoudre le problème de la recherche d'estimateurs d'erreur quadratique minimale. On se contentera de rechercher pour une famille de loi donnée $f(x ; \theta)$ l'estimateur ***sans biais*** de θ de ***variance minimale***. Il reste toutefois possible dans certains cas particuliers de trouver des estimateurs biaisés plus précis que le meilleur estimateur sans biais.

La recherche d'estimateurs sans biais de variance minimale est intimement liée à l'existence de statistiques exhaustives.

13.2 L'EXHAUSTIVITÉ

Dans un problème statistique où figure un paramètre θ inconnu, un échantillon apporte une certaine information sur ce paramètre (information qui serait différente pour un autre paramètre avec le même échantillon). Lorsque l'on résume cet échantillon par une statistique, il s'agit de ne pas perdre cette information ; une statistique qui conserve l'information sera qualifiée d'exhaustive.

Il convient de donner un sens précis à la notion d'information : une première approche consiste à remarquer qu'une variable aléatoire T ne peut nous renseigner sur la valeur d'un paramètre que dans la mesure où sa loi de probabilité dépend de ce paramètre ; si la variable T est une statistique relative à l'échantillon $(X_1, X_2, \ldots, X_n)$ et que la loi conditionnelle de $(X_1, X_2, \ldots, X_n)$ à T fixé ne dépend plus du paramètre θ, on peut dire alors, qu'une fois T connu, nous n'obtenons plus d'autre information de l'échantillon concernant θ et donc que T porte toute l'information disponible sur θ. Une deuxième approche consiste à définir mathématiquement une quantité d'information et à chercher dans quelles circonstances cette quantité se conserve lorsque les données sont résumées par une statistique.

13.2.1 Définition d'une statistique exhaustive

Soit un n-échantillon d'une variable aléatoire X.

On notera $L(x_1, x_2, \ldots, x_n ; \theta)$ soit la densité de $(X_1, X_2, \ldots, X_n)$ si X est absolument continue, soit la probabilité conjointe $P(X_1 = x_1 \cap \ldots \cap X_n = x_n)$ si X est discrète.

$L(\mathbf{x} ; \theta)$ considéré comme fonction de θ seul est appelé « ***vraisemblance*** » de θ (voir plus loin).

Soit T une statistique fonction de $X_1, X_2, \ldots, X_n$ de loi $g(t ; \theta)$ (densité dans le cas continu, $P(T = t)$ dans le cas discret).

DÉFINITION

T sera dite exhaustive si l'on a $L(\mathbf{x}, \theta) = g(t, \theta)h(\mathbf{x})$ (principe de factorisation) en d'autres termes si la densité conditionnelle de l'échantillon est indépendante du paramètre.

Ceci veut dire qu'une fois T connu, aucune valeur de l'échantillon ni aucune autre statistique ne nous apportera de renseignements supplémentaires sur θ.

Exemples :

- Loi normale, m connu σ inconnu :

$$L(\mathbf{x}, \sigma) = \frac{1}{\sigma^n (\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2\right]$$

Posons $T = \sum_{i=1}^{n}(X_i - m)^2$. On sait que T/σ^2 suit une loi de χ^2_n. La densité de T est

alors :

$$g(t, \sigma) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}\left(\frac{t}{\sigma^2}\right)^{\frac{n}{2}-1}\exp\left(-\frac{t}{2}\sigma^2\right)\frac{1}{\sigma^2}$$

$$= \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}\frac{t^{n/2-1}}{\sigma^n}\exp\left(-\frac{t}{2}\sigma^2\right)$$

d'où :

$$L(\mathbf{x}, \sigma) = g(t, \sigma)\frac{\Gamma(n/2)}{\pi^{n/2}\left[\sum_{i=1}^{n}(x_i - m)^2\right]^{n/2-1}} = g(t, \sigma)h(\mathbf{x})$$

$T = \sum_{i=1}^{n}(X_i - m)^2$ est donc exhaustif pour σ^2.

- Loi de Poisson, λ inconnu :

$$L(x_1 ; x_2 \dots x_n ; \lambda) = \prod_{i=1}^{n}\exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

$S = X_1 + X_2 + \cdots + X_n$ est exhaustive : S suit une loi $\mathscr{P}(n\lambda)$, d'où

$g(s ; \lambda) = \exp(-n\lambda)\dfrac{(n\lambda)^s}{s!}$ et :

$$\frac{L}{g} = \frac{s!}{n^s \prod x_i!}$$

Le principe de factorisation nous donne donc un moyen de reconnaître si une statistique est exhaustive, mais ne permet pas de la construire ou même de savoir s'il en existe une.

13.2.2 Lois permettant une statistique exhaustive

Le théorème suivant répond aux deux préoccupations précédentes :

Théorème de Darmois

Soit une variable aléatoire X dont le domaine de définition ne dépend pas de θ. Une condition nécessaire et suffisante pour que l'échantillon $(X_1, X_2, \ldots, X_n)$ admette une statistique exhaustive est que la forme de la densité soit :

$$f(x, \theta) = \exp[a(x)\alpha(\theta) + b(x) + \beta(\theta)] \quad (\textit{famille exponentielle})$$

Si la densité est de cette forme et si de plus l'application $x_1 \to \sum_{i=1}^{n} a(x_i)$ est bijective et continûment différentiable pour tout i, alors $T = \sum_{i=1}^{n} a(X_i)$ est une statistique exhaustive particulière.

Démonstration :

- Condition nécessaire : $T = \varphi(X_1, X_2, \ldots, X_n)$ est telle que :

$$L(\mathbf{x}, \theta) = g(t, \theta)\, h(\mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta)$$

On a :
$$\frac{\partial^2 \ln f(x_1, \theta)}{\partial\theta\, \partial x_i} = \frac{\partial^2 \ln g(t, \theta)}{\partial\theta\, \partial x_i} = \frac{\partial^2 \ln g(t, \theta)}{\partial\theta\, \partial t}\frac{\partial\varphi}{\partial x_i}$$

Posons :
$$k(\xi, \theta) = \frac{\partial \ln f(\xi, \theta)}{\partial\theta}$$

$\forall \xi, \theta \in \mathbb{R}$ il existe un point $\mathbf{x}$ de $\mathbb{R}^n$ avec $x_i = \xi x_j = \eta$. En ce point on a :

$$\frac{\partial k(\xi, \theta)/\partial\xi}{\partial k(\eta, \theta)/\partial\eta} = \frac{\partial\varphi(\mathbf{x})/\partial x_i}{\partial\varphi(\mathbf{x})/\partial x_j}$$

ce qui est indépendant de θ ; ceci n'est possible que si :

$$\frac{\partial k(x, \theta)}{\partial x} = u(x)v(\theta)$$

d'où en intégrant par rapport à x : $k(x, \theta) = a(x)v(\theta) + w(\theta)$, et en intégrant par rapport à θ : $\ln f(x, \theta) = a(x)\alpha(\theta) + \beta(\theta) + b(x)$.

- Condition suffisante : $L(x, \theta) = \exp\left[\alpha(\theta)\sum_{i=1}^{n} a(x_i) + \sum_{i=1}^{n} b(x_i) + n\beta(\theta)\right]$.

Posons $t = \Sigma a(x_i)$ et effectuons le changement de variable :

$$(x_1, x_2, \ldots, x_n) \longrightarrow (t, x_2, x_3, \ldots, x_n)$$

légitime si l'application est bijective $x_1 \rightarrow \sum_{i=1}^{n} a(x_i)$:

$$L'(t, x_2, \ldots, x_n) = \exp(t\alpha(\theta) + n\beta(\theta))\exp\left(\sum_{i=1}^{n} b(x_i)\right)\left|\frac{\partial t}{\partial x_1}\right|$$

car le jacobien de la transformation se réduit à $\partial t/\partial x_1$. Pour obtenir la densité $g(t, \theta)$ de t, il faut intégrer L' par rapport à $x_2, x_3, \ldots, x_n$ soit dans $\mathbb{R}^{n-1}$:

$$g(t, \theta) = \exp(t\alpha(\theta) + n\beta(\theta)) \cdot \int_{\mathbb{R}^{n-1}} \exp\left(\sum_{i=1}^{n} b(x_i)\right)\left|\frac{\partial t}{\partial x_1}\right| dx_2 dx_3 \ldots dx_n$$

il y a donc bien factorisation de $L(\mathbf{x}, \theta)$. ■

Ce théorème est un outil très puissant dans la recherche des statistiques exhaustives et l'on remarque que la plupart des lois usuelles, lois de Poisson, de Gauss, lois γ sont de la forme exponentielle.

■ **Exemple :** X suit une loi γ de paramètre inconnu :

$$f(x, \theta) = \frac{1}{\Gamma(\theta)}\exp(-x)x^{\theta-1}$$

$$\ln f(x, \theta) = -x + (\theta - 1)\ln x - \ln\Gamma(\theta)$$

La statistique exhaustive est $\sum_{i=1}^{n} \ln X_i = \ln\left(\prod_{i=1}^{n} X_i\right)$. ■

On peut remarquer que toute fonction injective d'une statistique exhaustive est encore exhaustive, ce qui indique que dans l'exemple précédent la moyenne géométrique des observations est exhaustive pour θ.

Une statistique exhaustive T, qui est fonction de toute statistique exhaustive, est dite exhaustive minimale.

Remarquons cependant que si le domaine de définition de X dépend de θ, le théorème de Darmois ne s'applique pas, ce qui n'empêche pas de trouver dans certains cas des statistiques exhaustives.

Ainsi si X suit une loi uniforme sur $[0 ; \theta]$, $T = \sup X_i$ est exhaustive pour θ.

En effet : $$L(\mathbf{x} ; \theta) = \left(\frac{1}{\theta}\right)^n \quad \text{et} \quad g(t ; \theta) = \frac{nt^{n-1}}{\theta^n}$$

car $P(T < t) = \left(\frac{t}{\theta}\right)^n$ il s'ensuit que $\frac{L}{g} = \frac{1}{nt^{n-1}}$ est indépendant de θ.

Autres exemples de statistiques exhaustives : le lecteur pourra vérifier les résultats suivants à titre d'exercice :

- loi de Bernoulli de paramètre p inconnu : $T = \sum_{i=1}^{n} X_i$ est exhaustif pour p ;
- loi de Laplace-Gauss : $N(m\,;\sigma)$:
  - si σ est connu, $T = \sum_{i=1}^{n} X_i$ est exhaustif pour m ;
  - si m est connu, $T = \sum_{i=1}^{n} (X_i - m)^2$ est exhaustif pour σ^2 ;
  - si m et σ sont tous deux inconnus, le couple $\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} (X_i - \overline{X})^2\right)$ ou $(\overline{X}, S^2)$ est exhaustif pour le couple (m, σ).
- loi exponentielle de densité $\frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$: $T = \sum_{i=1}^{n} X_i$ est exhaustif pour θ.

13.2.3 L'information de Fisher

DÉFINITION

On appelle quantité d'information de Fisher $I_n(\theta)$ apportée par un n-échantillon sur le paramètre θ la quantité suivante positive ou nulle (si elle existe) :

$$I_n(\theta) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]$$

Note : $L(\mathbf{X}, \theta)$ peut être considérée comme une variable aléatoire, car fonction de variable aléatoire :

$$L(X_1, X_2, \ldots, X_n\,; \theta) = \prod_{i=1}^{n} f(X_i\,; \theta)$$

THÉORÈME

Si le domaine de définition de X ne dépend pas de θ alors :

$$I_n(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right) \quad \textit{si cette quantité existe}$$

Démonstration : L étant une densité $\int_{\mathbb{R}^n} L(\mathbf{x}, \theta)d\mathbf{x} = 1$.

En dérivant les deux membres par rapport à θ et en remarquant que :

$$\frac{\partial L(\mathbf{x}, \theta)}{\partial \theta} = L(\mathbf{x}, \theta)\frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta}$$

il vient :
$$\int_{\mathbb{R}^n} \frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} L(\mathbf{x}, \theta) d\mathbf{x} = 0$$

ce qui prouve que la variable aléatoire $\frac{\partial \ln L(\mathbf{X}, \theta)}{\partial \theta}$ est centrée et que $I_n(\theta) = V\left(\frac{\partial \ln L}{\partial \theta}\right)$.

Dérivons une deuxième fois :

$$\int_{\mathbb{R}^n} \frac{\partial^2 \ln L(\mathbf{x}, \theta)}{\partial \theta} L(\mathbf{x}, \theta) d\mathbf{x} + \int_{\mathbb{R}^n} \frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} \frac{\partial L(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x} = 0$$

en utilisant à nouveau la remarque sur $\frac{\partial L(\mathbf{x}, \theta)}{\partial \theta}$, il vient :

$$\int_{\mathbb{R}^n} \frac{\partial^2 \ln L(\mathbf{x}, \theta)}{\partial \theta^2} L d\mathbf{x} + \int_{\mathbb{R}^n} \left(\frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta}\right)^2 L(\mathbf{x}, \theta) d\mathbf{x} = 0$$

ce qui démontre la proposition. ■

Remarque : L'utilisation de l'hypothèse du domaine indépendant de θ intervient lors de la dérivation sous le signe $\int$.

Un exemple de variable aléatoire à domaine non indépendant de θ est fourni par X de densité $\exp(-(x - \theta))$ si $x \geq \theta$; 0 sinon.

PROPRIÉTÉ DE $I_n(\theta)$

- Additivité. Si le domaine de définition ne dépend pas de θ on a :

$$I_n(\theta) = nI_1(\theta)$$

En effet les opérateurs espérance et dérivée seconde sont linéaires.

Ceci veut dire que chaque observation a la même importance, ce qui n'est pas le cas pour la loi uniforme sur [0, θ] où la plus grande observation est la plus intéressante.

- Précision. Soit X une variable aléatoire de Laplace-Gauss N(θ, σ) où σ est connu. On a $I_1(\theta) = 1/\sigma^2$: l'information apportée par une observation sur la moyenne est d'autant plus grande que la dispersion est petite.
- Dégradation de l'information. Montrons que l'information portée par une statistique est inférieure ou égale à celle apportée par l'échantillon. Soit T de densité $g(t, \theta)$ la statistique que l'on substitue à l'échantillon, on a :

$$L(\mathbf{x}, \theta) = g(t, \theta)h(\mathbf{x}, \theta | t)$$

où $h(\mathbf{x}, \theta \mid t)$ est la densité conditionnelle de l'échantillon. On a donc, en prenant l'espérance des dérivées secondes :

$$I_n(\theta) = I_T(\theta) - E\left(\frac{\partial^2 \ln h}{\partial \theta^2}\right)$$

le dernier terme est la quantité d'information conditionnelle $I_{n/T}(\theta)$ (ou information supplémentaire) ; elle est positive ou nulle, donc :

$$I_T(\theta) \leq I_n(\theta)$$

on voit donc que si T est exhaustive $I_T(\theta) = I_n(\theta)$ et que la réciproque est vraie si le domaine de X est indépendant de θ.

Remarque : On a supposé le domaine indépendant de θ car sinon on aurait dû écrire :

$$I_n(\theta) = I_t(\theta) + E\left[\left(\frac{\partial \ln h}{\partial \theta}\right)^2\right] + 2E\left[\frac{\partial \ln g}{\partial \theta}\frac{\partial \ln h}{\partial \theta}\right]$$

et on n'aurait pas pu conclure à une diminution de l'information à cause du signe inconnu du dernier terme.

Ce dernier terme peut laisser supposer, s'il est négatif et grand en valeur absolue, que $I_T(\theta) > I_n(\theta)$; jusqu'à présent aucun exemple d'augmentation de l'information n'a été découvert mais le problème reste entier.

13.2.4 Généralisation à plusieurs dimensions θ paramètre vectoriel $\in \mathbb{R}^s$

On consultera Fourgeaud, p. 216, pour un traitement complet. En résumé, on a, si le domaine ne dépend pas de $\boldsymbol{\theta}$:

La matrice de l'information $\mathcal{I}_n$ a pour terme général :

$$\mathcal{I}_{i,j} = \text{cov}\left[\frac{\partial \ln f(X, \theta)}{\partial \theta_i}, \frac{\partial \ln f(X, \theta)}{\partial \theta_j}\right]$$

c'est une matrice symétrique définie positive.

Soit $T_1, T_2, \ldots, T_s$ un système de s statistiques fonctionnellement indépendantes ; la notion de dégradation de l'information se généralise comme suit :

$$\mathcal{I}_n(\boldsymbol{\theta}) - \mathcal{I}_{T_1, T_2, \ldots, T_s}(\boldsymbol{\theta}) \text{ est définie positive}$$

On appelle système exhaustif un système de s statistiques fonctionnellement indépendantes, tel que :

$$L(x_1, x_2, \ldots, x_n ; \boldsymbol{\theta}) = g(t_1, t_2, \ldots, t_s ; \boldsymbol{\theta})h(x_1, x_2, \ldots, x_n)$$

et l'on a $\mathcal{I}_n(\boldsymbol{\theta}) - \mathcal{I}_{T_1, T_2, \ldots, T_s}(\boldsymbol{\theta}) = 0$ si et seulement si le système $(T_1, T_2, \ldots, T_s)$ est exhaustif.

Théorème de Darmois

Une condition nécessaire et suffisante pour qu'un n-échantillon admette un système résumé exhaustif est que :

$$\ln f(x, \theta) = \sum_{i=1}^{s} a_i(x)\alpha_i(\theta) + b(x) + \beta(\theta)$$

en particulier : $T_i = \sum_{j=1}^{n} a_i(X_j)$ $i = 1, 2, \ldots, s$ *est un systéme exhaustif*

13.3 L'ESTIMATION SANS BIAIS DE VARIANCE MINIMALE

13.3.1 Les résultats théoriques

On dispose pour résoudre ce problème d'une suite de quatre théorèmes qui montrent en définitive que l'estimateur de variance minimale est lié à l'existence d'une statistique exhaustive.

Théorème 1 Unicité

S'il existe un estimateur de θ *sans biais, de variance minimale, il est unique presque sûrement.*

Démonstration : Raisonnons par l'absurde et supposons qu'il existe deux estimateurs sans biais T_1 et T_2 de θ de variance minimale V.

Soit : $$T_3 = \frac{T_1 + T_2}{2}$$

T_3 est sans biais car : $$E(T_3) = \frac{E(T_1) + E(T_2)}{2} = \frac{\theta + \theta}{2}$$

et : $$V(T_3) = \frac{1}{4}[V(T_1) + V(T_2) + 2\rho\sigma_{T_1}\sigma_{T_2}]$$

où ρ est le coefficient de corrélation linéaire entre T_1 et T_2. Puisque $V(T_1) = V(T_2) = V$ il vient $V(T_3) = \frac{V}{2}(1 + \rho)$. Si $\rho < 1$ on a $V(T_3) < V$ ce qui est impossible, donc $\rho = 1$. C'est-à-dire $T_1 - E(T_1) = \lambda(T_2 - E(T_2))$ avec $\lambda > 0$. Comme $V(T_1) = V(T_2)$ il vient $\lambda = 1$ et puisque $E(T_1) = E(T_2) = \theta$ on a $T_1 = T_2$ (ps).

Théorème 2 : Rao-Blackwell

Soit T un estimateur quelconque sans biais de θ *et U une statistique exhaustive pour* θ*. Alors* $T^* = E(T \mid U)$ *est un estimateur sans biais de* θ *au moins aussi bon que T.*

Démonstration :

- **T^* est un estimateur de θ.** Cette proposition est non triviale car il faut montrer que T^* dépend seulement des X_i et non de θ.

Puisque U est exhaustive, la densité conditionnelle de l'échantillon sachant U ne dépend pas de θ et $E(T|U) = \int_{\mathbb{R}^n} tL(\mathbf{x},\theta|u)d\mathbf{x}$ ne dépend donc pas de θ mais des x seuls.

- **T^* est sans biais.** D'après le théorème de l'espérance totale :

$$E(T^*) = E[E(T|U)] = E(T) = \theta$$

- **T^* est au moins aussi bon que T.** D'après le théorème de la variance totale :

$$V(T) = V(E(T|U)) + E(V(T|U))$$
$$V(T) = V(T^*) + E(V(T|U))$$

Comme $E(V(T|U))$ est positif ou nul on a $V(T) \geq V(T^*)$.

- De plus si $E(V(T|U)) = 0$ c'est que presque sûrement $T = f(U)$, il y a relation fonctionnelle entre T et U.

Ce théorème fournit une méthode pour améliorer un estimateur sans biais donné.

THÉORÈME 3

S'il existe une statistique exhaustive U, alors l'estimateur T sans biais de θ de variance minimale (unique d'après le théorème 1) ne dépend que de U.

C'est un corollaire du théorème 2. On ne peut améliorer T par la méthode de Rao-Blackwell puisque T est de variance minimale. Donc $V(T^*) = V(T)$ et $T = f(U)$.

Cependant, comme il peut exister plusieurs estimateurs sans biais de θ fonction de U, on n'est pas sûr que l'estimateur T^* obtenu par la méthode de Rao-Blackwell soit le meilleur, il faut alors introduire la notion de statistique complète.

DÉFINITION

On dit qu'une statistique U est complète pour une famille de lois de probabilités $f(x, \theta)$ si $E[h(U)] = 0 \; \forall \theta \Rightarrow h = 0$ ps.

On montre en particulier que la statistique exhaustive des familles exponentielles est complète.

Ainsi par exemple pour une loi de Poisson $\mathcal{P}(\lambda)$ où λ est inconnu $S = \sum_{i=1}^{n} X_i$ est complète.

En effet :

$$E[h(S)] = \sum_{s=0}^{\infty} h(s) \exp(-n\lambda) \frac{(n\lambda)^s}{s!}$$
$$= \exp(-n\lambda) \sum_{s=0}^{\infty} h(s) \frac{(n\lambda)^s}{s!}$$

La série $\sum_{s=0}^{\infty} \frac{h(s)n^s}{s!}\lambda^s$ ne peut être nulle $\forall\lambda$ que si elle est nulle terme à terme donc si $h(s) = 0 \quad \forall s \in \mathbb{N}$.

THÉORÈME 4 : LEHMANN-SCHEFFÉ

Si T^* est un estimateur sans biais de θ dépendant d'une statistique exhaustive complète U alors T^* est l'unique estimateur sans biais de variance minimale de θ. En particulier si l'on dispose déjà de T estimateur sans biais de θ, $T^* = E(T \mid U)$.

En effet l'estimateur de variance minimale est unique et dépend de U, d'autre part U étant complète il n'existe qu'un seul estimateur sans biais dépendant de U (soit $T_1 = f(U)$ et $T_2 = g(U)$ $\quad E(T_1) - E(T_2) = 0 \quad \forall\theta \Longrightarrow f = g$ ps) l'estimateur obtenu est donc nécessairement le meilleur.

En conclusion si l'on dispose d'un estimateur sans biais fonction d'une statistique exhaustive complète, c'est le meilleur estimateur possible.

13.3.2 Exemple

Le nombre de demandes hebdomadaires d'un certain produit est une variable X qui suit une loi de Poisson $\mathcal{P}(\lambda)$ où λ est inconnu. On cherche à évaluer la probabilité que X soit nul. On note $X_1, X_2, \ldots, X_n$ les observations de X pendant n semaines.

Le paramètre à estimer est $\exp(-\lambda) = P(X = 0)$.

Une première méthode consiste à compter le nombre de fois K où l'on a observé $X = 0$ et à estimer $P(X = 0)$ par la fréquence K/n.

On a bien sûr :

$$E\left(\frac{K}{n}\right) = \exp(-\lambda) \qquad V\left(\frac{K}{n}\right) = \frac{\exp(-\lambda)(1 - \exp(-\lambda))}{n}$$
$$= \exp(-2\lambda)\left(\frac{\exp(\lambda) - 1}{n}\right)$$

K/n est sans biais, convergent, mais ne tient pas compte du fait que X suit une loi de Poisson. Il ne peut donc être optimal, ce qui se vérifie en remarquant que K/n n'est pas une fonction de $S = \sum_{i=1}^{n} X_i$ qui est une statistique exhaustive complète pour λ.

$\overline{X}$ est l'estimateur de variance minimale de λ, mais $\exp(-\overline{X})$ est biaisé pour $\exp(-\lambda)$.

L'estimateur sans biais de variance minimale T de $\exp(-\lambda)$ peut être obtenu en améliorant K/n par l'application du théorème de Rao-Blackwell :

Introduisons les variables de Bernoulli : $Y_1, Y_2, \ldots, Y_n$:

$$\begin{cases} Y_i = 1 & \text{si } X_i = 0 & \text{d'où } P(Y_i = 1) = \exp(-\lambda) \\ Y_i = 0 & \text{si } X_i > 1 & \text{d'où } P(Y_i = 0) = 1 - \exp(-\lambda) \end{cases}$$

$$K = \sum_{i=1}^{n} Y_i$$

On a : $$T = E\left[\frac{K}{n}\Big/ S\right] = \frac{1}{n}E[K/S] = E[Y_1/S]$$

Or : $$E[Y_1/S = s] = P(Y_1 = 1/S = s) = P(X_1 = 0/S = s)$$

D'après la formule de Bayes :

$$P(X_1 = 0/S = s) = \frac{P(S = s/X_1 = 0)P(X_1 = 0)}{P(S = s)}$$

S suit une loi de Poisson $\mathcal{P}(n\lambda)$. La loi de S, sachant que $X_1 = 0$, est une loi de Poisson $\mathcal{P}((n-1)\lambda)$ car alors $S/(X_1 = 0) = X_2 + X_3 + \cdots + X_n$.

D'où : $$P(X_1 = 0/S = s) = \frac{[\exp(-(n-1)\lambda)]\dfrac{((n-1)\lambda)^s}{s!}\exp(-\lambda)}{[\exp(-n\lambda)]\dfrac{(n\lambda)^s}{s!}}$$

$$= \left(\frac{n-1}{n}\right)^s = \left(1 - \frac{1}{n}\right)^{n\bar{x}}$$

Donc : $$T = \left(1 - \frac{1}{n}\right)^{n\bar{X}}$$

Un calcul laissé au soin du lecteur montre que $V(T) = \exp(-2\lambda)\left(\exp\left(\frac{\lambda}{n}\right) - 1\right)$; on a donc bien $V(T) < V\left(\frac{K}{n}\right)$ car :

$$V(T) = \exp(-2\lambda)\left[\frac{\lambda}{n} + \frac{\lambda^2}{2n^2} + \cdots + \frac{\lambda^k}{k!n^k} + \cdots\right]$$

$$V\left(\frac{K}{n}\right) = \exp(-2\lambda)\left[\frac{\lambda}{n} + \frac{\lambda^2}{2n} + \cdots + \frac{\lambda^k}{k!n} + \cdots\right]$$

13.3.3 Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR)

Le résultat suivant nous indique que la variance d'un estimateur ne peut être inférieure à une certaine borne, qui dépend de la quantité d'information de Fisher apportée par l'échantillon sur le paramètre θ.

Si le domaine de définition de X ne dépend pas de θ, on a pour tout estimateur T sans biais de θ :

$$V(T) \geq \frac{1}{I_n(\theta)}$$

et si T est un estimateur sans biais de $h(\theta)$:

$$\boxed{V(T) \geq \frac{[h'(\theta)]^2}{I_n(\theta)}}$$

Démonstration : Considérons :

$$\text{cov}\left(T, \frac{\partial \ln L}{\partial \theta}\right) = E\left(T \frac{\partial \ln L}{\partial \theta}\right)$$

puisque $\frac{\partial \ln L}{\partial \theta}$ est centrée. Donc :

$$\text{cov}\left(T, \frac{\partial \ln L}{\partial \theta}\right) = \int t \frac{\partial \ln L}{\partial \theta} L \mathrm{d}\mathbf{x} = \int t \frac{\partial L}{\partial \theta} \mathrm{d}\mathbf{x}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\theta} \int tL \mathrm{d}\mathbf{x} = \frac{\mathrm{d}}{\mathrm{d}\theta} E(T) = h'(\theta)$$

D'autre part l'inégalité de Schwarz donne :

$$\left[\text{cov}\left(T, \frac{\partial \ln L}{\partial \theta}\right)\right]^2 \leq V(T)\, V\left(\frac{\partial \ln L}{\partial \theta}\right)$$

c'est-à-dire : $[h'(\theta)]^2 \leq V(T)\, I_n(\theta)$ c.q.f.d.

La question se pose de savoir si l'on peut atteindre la borne minimale de la variance ; un tel estimateur sera qualifié d'efficace.

L'efficacité n'est donc définie que dans les conditions de régularité suivantes qui sont celles de FDCR :

a) Le domaine de définition E_θ est indépendant de θ.

b) $\frac{\partial L}{\partial \theta}$ existe et est continue par rapport à θ.

c) $I_n(\theta)$ est finie.

d) $\frac{\partial L}{\partial \theta}$, $T \frac{\partial L}{\partial \theta}$ sont intégrables par rapport à θ.

Dire que T est efficace c'est dire que sous ces conditions :

$$V(T) = \frac{[h'(\theta)]^2}{I_n(\theta)} \quad \forall \theta \in \Theta$$

T est donc un estimateur sans biais de variance minimale de $h(\theta)$.

On a alors le résultat suivant :

THÉORÈME SUR L'EFFICACITÉ

- *La borne de Cramer-Rao ne peut être atteinte que si la loi de X est de forme exponentielle :*

$$\ln f(x,\theta) = a(x)\alpha(\theta) + b(x) + \beta(\theta)$$

car T est nécessairement exhaustif pour θ.

- *Si la loi de X est bien de la forme précédente, il n'existe (à une transformation linéaire près) qu'une seule fonction* $h(\theta)$ *du paramètre qui puisse être estimée efficacement : c'est* $h(\theta) = -\dfrac{\beta'(\theta)}{\alpha'(\theta)}$.

L'estimateur de $h(\theta)$ *est alors :*

$$T = \frac{1}{n}\sum_{i=1}^{i=n} a(X_i)$$

La variance minimale est :

$$V(T) = -\frac{1}{n\alpha'(\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\beta'(\theta)}{\alpha'(\theta)}\right) = \frac{h'(\theta)}{n\alpha'(\theta)}$$

Démonstration :

- T est exhaustif si T est efficace de $h(\theta)$.

Comme E_θ ne dépend pas de θ on a toujours :

$$I_T(\theta) \leq I_n(\theta)$$

Une conséquence de FDCR est que :

$$V(T) \geq \frac{[h'(\theta)]^2}{I_T(\theta)}$$

Si T est efficace on a :

$$V(T) = \frac{[h'(\theta)]^2}{I_n(\theta)}$$

donc :

$$I_n(\theta) \leq I_T(\theta)$$

donc $I_n(\theta) = I_T(\theta)$. T est donc exhaustive.

D'après le théorème de Darmois on a alors :

$$\ln f(x, \theta) = a(x)\alpha(\theta) + \beta(\theta) + b(x)$$

- Si T est efficace pour $h(\theta)$ et si $\frac{1}{n}\sum_{i=1}^{n} a(X_i)$ est exhaustif alors :

$$h(\theta) = \frac{\beta'(\theta)}{\alpha'(\theta)} \qquad \text{et} \qquad T = \frac{1}{n}\sum_{i=1}^{n} a(X_i)$$

L'inégalité de FDCR étant une inégalité de Schwarz, l'égalité n'est réalisée que s'il y a colinéarité pour presque tout θ, c'est-à-dire :

$$\frac{\partial \ln L}{\partial \theta} = \lambda(\theta)[T - h(\theta)]$$

Or, si $L = \prod_{i=1}^{n} \exp(a(x_i)\alpha(\theta) + \beta(\theta) + b(x_i))$ on doit avoir :

$$\ln L = \alpha(\theta)\sum_{i=1}^{n} a(x_i) + n\beta(\theta) + \sum_{i=1}^{n} b(x_i)$$

et : $$\frac{\partial \ln L}{\partial \theta} = \alpha'(\theta)\sum_{i=1}^{n} a(x_i) + n\beta'(\theta) = n\alpha'(\theta)\left[\frac{1}{n}\sum_{i=1}^{n} a(x_i) + \frac{n\beta'(\theta)}{\alpha'(\theta)}\right],$$

En identifiant les deux expressions de $\frac{\partial \ln L}{\partial \theta}$ on obtient :

$$T = \frac{1}{n}\sum_{i=1}^{n} a(X_i) \qquad \text{et} \qquad h(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$$

$h(\theta)$ et T sont donc uniques par construction à une transformation affine près.

La réciproque est alors immédiate :

Si la loi est de la famille exponentielle, la statistique exhaustive $T = \frac{1}{n}\sum_{i=1}^{n} a(X_i)$ est efficace pour $-\frac{\beta'(\theta)}{\alpha'(\theta)}$.

- Calcul de la variance $V(T)$:

De $I_n(\theta) = n^2\,\alpha'(\theta)^2 V(T)$ et $V(T) = \frac{[h'(\theta)]^2}{I_n(\theta)}$ on déduit :

$$V(T) = \frac{1}{n}\left|\frac{h'(\theta)}{\alpha'(\theta)}\right|$$

On peut montrer que $V(T) = \frac{1}{n}\frac{h'(\theta)}{\alpha'(\theta)}$ ■

Le théorème qui vient d'être démontré montre qu'on ne peut estimer efficacement qu'une seule fonction $h(\theta)$ qui peut ne pas être intéressante.

Exemple 1. Estimation du paramètre θ d'une loi γ_0 :

$$\ln f(x, \theta) = (\theta - 1)\ln x - x - \ln \Gamma(\theta)$$

Si l'on prend $T = \frac{1}{n}\sum_{i=1}^{n} \ln X_i = \ln\left(\prod_{i=1}^{n} X_i\right)^{1/n}$ comme estimateur, on voit que l'on estime efficacement $h(\theta) = \frac{d}{d\theta}\ln(\Gamma(\theta))$.

Exemple 2. Dans une loi $N(m, \sigma)$, si m est connu σ^2 est le seul paramètre que l'on peut estimer efficacement et ceci par $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$. L'estimateur $\sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\sqrt{T}$ est sans biais pour σ, de variance minimale car T est exhaustive, mais n'est pas efficace au sens de la borne de FDCR. Si m est inconnu l'estimateur $\sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}S$ est sans biais et de variance minimale pour σ. En pratique on utilisera $S^* = \sqrt{\frac{n}{n-1}}S$ qui est très légèrement biaisé (voir § 12.2.3).

Remarque : Si X ne suit pas une loi $N(m, \sigma)$ on ne peut donner d'expression universelle d'un estimateur sans biais de σ.

La recherche de statistiques exhaustives peut ne pas aboutir. on possède cependant une méthode d'obtention de bons estimateurs.

13.4 LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE (MV)

Cette méthode consiste, étant donné un échantillon de valeurs $x_1, x_2, \ldots, x_n$ à prendre comme estimation de θ la valeur de θ qui rend maximale la vraisemblance :

$$L(x_1, x_2, \ldots, x_n ; \theta).$$

En pratique on prend comme estimation de θ une solution de l'équation $\frac{\partial}{\partial\theta}\ln L(X ; \theta) = 0$, dite "équation de la vraisemblance".

Intuitivement, puisque L représente une densité de probabilité, cela revient à supposer que l'événement qui s'est produit était le plus « probable ».

Les justifications mathématiques sont les suivantes :

Propriété 1

S'il existe une statistique exhaustive U, alors l'estimateur du MV en dépend.

En effet $L(\mathbf{x}, \theta) = g(u, \theta)h(\mathbf{x})$ et résoudre $\frac{\partial \ln L}{\partial \theta} = 0$ revient à résoudre $\frac{\partial \ln g}{\partial \theta} = 0$ donc $\hat{\theta} = f(u)$.

Si $\hat{\theta}$ est sans biais, ce qui n'est pas toujours réalisé, $\hat{\theta}$ sera la meilleure estimation possible de θ si les conditions des théorèmes précédents sont réalisées.

Propriété 2. Invariance fonctionnelle

Si $\hat{\theta}$ est l'estimateur du MV de θ, $f(\hat{\theta})$ est l'estimateur du MV de $f(\theta)$.

La démonstration est élémentaire si f est bijective, plus délicate dans le cas général.

S'il n'existe pas de statistique exhaustive U on a les propriétés asymptotiques suivantes.

Propriété 3 (admise)

Il existe une suite de valeurs $\hat{\theta}_n$ racines de l'équation de la vraisemblance qui converge presque sûrement vers θ si $\rightarrow \infty$. De plus $\exists N$ tel que $n > N$ entraîne que $\hat{\theta}_n$ réalise effectivement un maximum pour L.

Propriété 4 (admise)

$$\frac{\hat{\theta}_n - \theta}{\frac{1}{\sqrt{I_n(\theta)}}} \xrightarrow{\mathcal{L}} N(0, 1)$$

On peut donc affirmer, avec certaines réserves, qu'asymptotiquement $V(\hat{\theta}_n) \rightarrow \frac{1}{I_n(\theta)}$, donc que $\hat{\theta}_n$ est asymptotiquement efficace.

Remarques : L'équation de la vraisemblance n'a pas nécessairement une racine unique. De plus cette méthode n'est valable utilement que pour de grands échantillons, à cause de ses propriétés asymptotiques, s'il n'existe pas de statistique exhaustive U.

Exemple : Estimation du paramètre de la loi de Weibull standard :

$$F(x) = \exp(-x^{\theta})$$
$$f(x ; \theta) = \theta x^{\theta-1}\exp(-x^{\theta})$$

Le domaine de définition ne dépend pas de θ, mais la loi n'est pas de la forme de Darmois, à cause du terme en x^{θ}. Appliquons la méthode du maximum de vraisemblance :

$$L(\mathbf{x} ; \theta) = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}\exp(-\sum_{i=1}^{n} x_i^{\theta})$$

$$\ln L(\mathbf{x} ; \theta) = n \ln \theta + (\theta - 1)\sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} x_i^{\theta}$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} x_i^{\theta} \ln x_i$$

$\hat{\theta}$ est donc solution de l'équation :

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} (x_i^{\hat{\theta}} - 1)\ln x_i}$$

Cette équation non linéaire ne peut se résoudre que numériquement par approximations successives et on ne peut donc pas obtenir de forme explicite pour l'estimateur de θ. ■

Extension à plusieurs paramètres $\theta_1, \theta_2, \ldots, \theta_p$:

La méthode consiste alors à résoudre le système d'équations simultanées :

$$\frac{\partial \ln L}{\partial \theta_j} = 0 \qquad j = 1, 2, \ldots, p$$

Les propriétés de convergence et d'invariance fonctionnelle s'étendent sans difficulté et on a également la propriété de normalité asymptotique suivante $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)$ a, quand n tend vers l'infini, une distribution qui tend vers une loi gaussienne à p dimensions de vecteur espérance $\theta_1, \theta_2, \ldots, \theta_p$ et dont la matrice de variance est l'inverse de la matrice d'information de Fisher.

Plus précisément si le domaine de définition ne dépend pas des paramètres à estimer : Σ^{-1} a pour terme général $-E\left[\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right]$.

13.5 L'ESTIMATION PAR INTERVALLES (LES FOURCHETTES D'UNE ESTIMATION)

Il est souvent plus réaliste et plus intéressant de fournir un renseignement du type $a < \theta < b$ plutôt que d'écrire sèchement $\hat{\theta} = c$.

Fournir un tel intervalle $[a, b]$ s'appelle donner une estimation par intervalle de θ ou estimation ensembliste.

13.5.1 Principe

La méthode des intervalles de confiance est la suivante :

Soit T un estimateur de θ, (on prendra évidemment le meilleur estimateur possible), dont on connaît la loi de probabilité pour chaque valeur de θ.

Étant donné une valeur θ_0 de θ, on détermine un intervalle de probabilité de niveau $1 - \alpha$ pour T, c'est-à-dire deux bornes t_1 et t_2 telles que :

$$P(t_1 < T < t_2 | \theta = \theta_0) = 1 - \alpha$$

Ces bornes dépendent évidemment de θ_0.

On choisit dans la plupart des cas un intervalle de probabilité à risques symétriques $\alpha/2$ et $\alpha/2$.

On adopte alors la règle de décision suivante : soit t la valeur observée de T :

- si $t \in [t_1, t_2]$ on conserve θ_0 comme valeur possible de θ ;
- si $t \notin [t_1, t_2]$ on élimine θ_0.

On répète cette opération pour toutes les valeurs de θ.

On peut traduire graphiquement cette méthode dans un plan (θ ; T) où l'on trace $t_1(\theta)$ $t_2(\theta)$ (fig. 13.2).

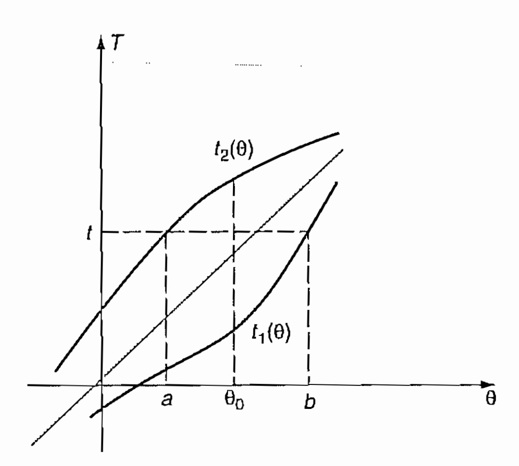


FIGURE 13.2

On lit donc selon une verticale les intervalles de probabilité et, selon l'horizontale issue de t, l'intervalle de confiance [a, b] de niveau $1 - \alpha$ (coefficient de confiance).

$$\begin{cases} a = t_2^{-1}(t) \\ b = t_1^{-1}(t) \end{cases}$$

[a, b] est un intervalle aléatoire car il dépend de t.

Si l'on augmente $1 - \alpha$, on augmente la longueur de l'intervalle de probabilité, donc les courbes s'écartent.

Si n augmente, comme T est supposé convergent, $V(T)$ diminue, donc $[t_1, t_2]$ diminue et les courbes se rapprochent de la première bissectrice.

13.5.2 Espérance d'une variable normale

13.5.2.1 σ est connu

$\overline{X}$ est le meilleur estimateur de m et $\overline{X}$ suit une loi $LG\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

L'intervalle de probabilité de $\overline{X}$ à $1 - \alpha$ est :

$$m - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < m + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

d'où l'intervalle de confiance :

$$\bar{x} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < m < \bar{x} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

si $1 - \alpha = 0.95$ on a $u_{\alpha/2} = 1.96$.

13.5.2.2 σ est inconnu

On utilise le fait que $T = \frac{\overline{X} - m}{S}\sqrt{n - 1}$ suit une loi de Student à $(n - 1)$ degrés de liberté.

L'intervalle de probabilité pour t est :

$$-t_{\alpha/2} < \frac{\overline{X} - m}{S}\sqrt{n - 1} < t_{\alpha/2}$$

d'où l'intervalle de confiance :

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n - 1}} < m < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n - 1}}$$

ou bien :

$$\boxed{\bar{x} - t_{\alpha/2}\frac{s^*}{\sqrt{n}} < m < \bar{x} + t_{\alpha/2}\frac{s^*}{\sqrt{n}}}$$

Le théorème central-limite a pour conséquence que les intervalles précédents sont valables pour estimer m d'une loi quelconque que n est assez grand.

13.5.3 Variance d'une loi normale

13.5.3.1 m est connu

$T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ est le meilleur estimateur de σ^2 et $\frac{nT}{\sigma^2}$ suit un χ^2_n comme somme de n carrés de $LG(0, 1)$ indépendantes.

Soit k_1 et k_2 les bornes de l'intervalle de probabilité d'un χ^2_n (fig. 13.3) :

$$P\left(k_1 < \frac{nT}{\sigma^2} < k_2\right) = 1 - \alpha$$

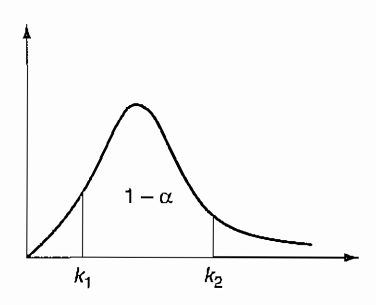


FIGURE 13.3

L'intervalle de confiance est : $\dfrac{nt}{k_2} < \sigma^2 < \dfrac{nt}{k_1}$

13.5.3.2 m est inconnu

On utilise $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ et on sait que $\dfrac{nS^2}{\sigma^2}$ suit χ^2_{n-1}, soit l_1 et l_2 les bornes de l'intervalle de probabilité :

$$P\left(l_1 < \frac{nS^2}{\sigma^2} < l_2\right) = 1 - \alpha$$

On a alors : $\boxed{\dfrac{ns^2}{l_2} < \sigma^2 < \dfrac{ns^2}{l_1}}$

Exemple : $n = 30$; $s^2 = 12$; $1 - \alpha = 0.90$; $8.46 < \sigma^2 < 20.33$ d'où $2.91 < \sigma < 4.51$.

Note importante : Ces formules ne sont valables que si x suit une loi normale.

13.5.4 Intervalle de confiance pour une proportion p

Étant donné une population infinie (ou finie si le tirage s'effectue avec remise) où une proportion p des individus possède un certain caractère, il s'agit de trouver un intervalle de confiance pour p à partir de f, proportion trouvée dans un échantillon de taille n.

On sait que nf suit une loi binomiale $\mathcal{B}(n, p)$; si n est faible on utilisera les tables de la loi binomiale ou l'abaque (voir Table A3. bis).

Si n est grand on sait que $nF \sim N(np\,;\sqrt{np(1-p)})$ donc que :

$$F \sim N\left(p\,;\sqrt{\frac{p(1-p)}{n}}\right)$$

L'intervalle de probabilité symétrique est :

$$p - u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < F < p + u_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Posons $u_{\alpha/2} = k$ pour simplifier les notations.

Les bornes de l'intervalle de probabilité sont données par :

$$y = p \pm k\sqrt{\frac{p(1-p)}{n}}$$

soit :
$$(y-p)^2 = \frac{k^2 p(1-p)}{n}$$

ou :
$$y^2 + p^2\left(1 + \frac{k^2}{n}\right) - 2py - \frac{k^2 p}{n} = 0$$

Équation d'une ellipse passant par l'origine et le point (1, 1), points pour lesquels elle a une tangente verticale (fig. 13.4).

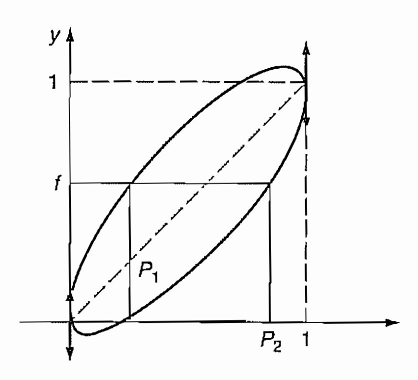


FIGURE 13.4

Les parties de l'ellipse extérieure au carré unité sont sans signification ; elles correspondent aux zones où l'approximation normale n'est pas valable.

Étant donné une valeur f observée, l'intervalle de confiance s'obtient en résolvant en p l'équation :

$$f^2 + p^2\left(1 + \frac{k^2}{n}\right) - 2pf - \frac{k^2}{n}p = 0$$

ou : $$p^2\left(1+\frac{k^2}{n}\right)-p\left(\frac{k^2}{n}+2f\right)+f^2=0$$

Résolvons-la complètement :

$$\Delta=\left(\frac{k^2}{n}+2f\right)^2-4\left(1+\frac{k^2}{n}\right)f^2=\frac{k^4}{n^2}+4f\frac{k^2}{n}-4f^2\frac{k^2}{n}$$

d'où : $$p=\frac{\left(2f+\frac{k^2}{n}\right)\pm\sqrt{\frac{k^4}{n^2}+4f\frac{k^2}{n}-4f^2\frac{k^2}{n}}}{2\left(1+\frac{k^2}{n}\right)}$$

formule encombrante mais dont on peut trouver une approximation en considérant que n est grand et en faisant un développement limité au premier ordre en $(1/n)$; le premier terme $\frac{2f+\frac{k^2}{n}}{2\left(1+\frac{k^2}{n}\right)^2}\sim f+0\left(\frac{1}{n^2}\right)$, le second se réduit en simplifiant par n^2 :

$$\sqrt{\frac{k^4+4fnk^2-4f^2nk^2}{4(n+k^2)^2}}=\sqrt{\frac{k^4+4fnk^2-4f^2nk^2}{4n^2+8k^2n+4k^4}}$$

ce radical est équivalent au suivant (en écrivant que chaque terme est équivalent à celui du plus haut degré en n) :

$$\sqrt{\frac{fnk^2-f^2nk^2}{n^2}}=k\sqrt{\frac{f(1-f)}{n}}$$

donc, on a si n est grand, l'expression approchée suivante pour l'intervalle de confiance :

$$\boxed{f-u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}<p<f+u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}}$$

Exemple : $n=400$; $f=36\ \%$; $1-\alpha=0.95$. On a $0.31<p<0.41$.

Application : Détermination de la taille d'un échantillon en fonction de la précision souhaitée.

Supposons que l'on désire connaître p avec une incertitude $\pm\Delta p$ pour un niveau de confiance donné $1-\alpha$ à risques symétriques. La formule précédente nous indique que :

$$\Delta p=u_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$$

soit : $$n=\frac{(u_{\alpha/2})^2f(1-f)}{(\Delta p)^2}$$

Si f est inconnu on obtient une majoration de n en posant $f = 1/2$ (cas le plus défavorable pour un sondage). D'où la relation :

$$n \leq \frac{(u_\alpha)^2}{4(\Delta p)^2}$$

Dans le cas d'un intervalle à 95 %, $u_{\alpha/2} = 1.96$ # 2, ce qui donne la formule approchée :

$$n_{\max} = \frac{1}{(\Delta p)^2}$$

on a les valeurs approchées suivantes de n :

Δp \ $1-\alpha$	0.90	0.95	0.98
0.01	6 760	9 600	13 530
0.02	1 700	2 400	3 380
0.05	270	380	540

Remarque : Les formules précédentes sont souvent abusivement utilisées (en particulier dans les médias) pour expliquer les marges d'erreur des sondages d'opinion. On ne peut en réalité pas les appliquer aux sondages d'opinion, sauf à la rigueur pour donner une borne supérieure de l'erreur d'échantillonnage : en effet, les sondages d'opinion ne sont pas effectués selon la méthode du tirage aléatoire simple équiprobable dans la totalité de la population, mais selon des techniques bien plus complexes (stratification, quotas etc. *cf.* chapitre 20) qui conduisent à diminuer la marge d'erreur du sondage simple en utilisant des informations auxiliaires. Signalons enfin que ces calculs de variance ne servent qu'à calculer l'erreur dûe au tirage aléatoire des observations ; l'échantillonnage n'est qu'une des sources d'erreur, pas toujours la plus importante, laquelle s'ajoute bien d'autres types d'erreurs : non-réponse ou refus, dissimulation, incompréhension des questions etc.

13.5.5 Intervalle de confiance pour le paramètre λ d'une loi de Poisson

Soit $\bar{x}$ la moyenne d'un n-échantillon d'une variable $\mathcal{P}(\lambda)$.

Comme on sait que $P(X \leq k) = P(\chi^2_{2(k+1)} > 2\lambda)$, on en déduit l'intervalle de confiance pour λ à risques symétriques de niveau $1 - \alpha$:

$$\boxed{\frac{1}{2n}\chi^2_{2n\bar{x}\,;\,\alpha/2} \leq \lambda \leq \frac{1}{2n}\chi^2_{2(n\bar{x}+1)\,;\,1-\frac{\alpha}{2}}}$$

où $\chi^2_{c\,;\,\alpha}$ est le quantile d'ordre α d'un χ^2 à c degrés de liberté.

Exemple : $n = 15$; $n\bar{x} = \sum_{i=1}^{n} x_i = 20$; $\alpha = 0.1$.

$$\frac{1}{30}\chi^2_{40\,;\,5\%} \leq \lambda \leq \frac{1}{30}\chi^2_{42\,;\,95\%}$$

$$\frac{26.5}{30} \leq \lambda \leq \frac{58.1}{30} \qquad \text{soit } 0.88 \leq \lambda \leq 1.94$$

Pour les grandes valeurs de n, lorsque $2n\bar{x}$ dépasse les possibilités des tables de χ^2, on utilisera une des approximations normales de la loi du χ^2. Si l'on utilise l'approximation de Wilson et Hilferty, qui est de loin la plus précise, on a :

$$\bar{x}\left(1 - \frac{u}{3\sqrt{n\bar{x}}} - \frac{1}{9n\bar{x}}\right)^3 \leq \lambda \leq \left(\bar{x} + \frac{1}{n}\right)\left(\frac{u}{3\sqrt{n\bar{x}+1}} + 1 - \frac{1}{9(n\bar{x}+1)}\right)^3$$

13.5.6 Ellipsoïde de confiance pour la moyenne d'une loi de Gauss multidimensionnelle

On a vu au chapitre 12 paragr. 12.3 que le centre de gravité d'un n-échantillon suivant une loi $N_p(\boldsymbol{\mu}\,;\,\boldsymbol{\Sigma})$ était tel que si $\boldsymbol{\Sigma}$ est connu :

$$n(\mathbf{g} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{g} - \boldsymbol{\mu}) = \chi^2_p$$

ou si $\boldsymbol{\Sigma}$ est inconnu :

$$\frac{(n-p)}{p}(\mathbf{g} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{g} - \boldsymbol{\mu}) = F(p\,;\,n-p)$$

On peut donc en déduire des zones de confiance ellipsoïdales de $\boldsymbol{\mu}$ autour de $\mathbf{g}$ définies par :

$$\boxed{(\mathbf{g} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{g} - \boldsymbol{\mu}) \leq \frac{p}{n-p}F_{1-\alpha}(p\,;\,n-p)}$$

Pour $p = 2$ on a des zones elliptiques dans le plan. Lorsque n est très grand, toujours pour $p = 2$, l'ellipse à 95 % a pour équation approximative :

$$(\mathbf{g} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{g} - \boldsymbol{\mu}) \simeq \frac{6}{n}$$

Les résultats précédents s'appliquent en particulier pour les estimateurs du maximum de vraisemblance car ils sont asymptotiquement normaux.

La figure suivante donne l'ellipse de confiance à 95 % pour la position simultanée des moyennes de deux variables dans un échantillon de 24 observations (prix et superficie d'appartements parisiens*). La forme elliptique est ici très accentuée car le coefficient de corrélation entre les deux variables est élevé r = 0.9733

* Les données sont présentées au chapitre 16, § 16.4.1.

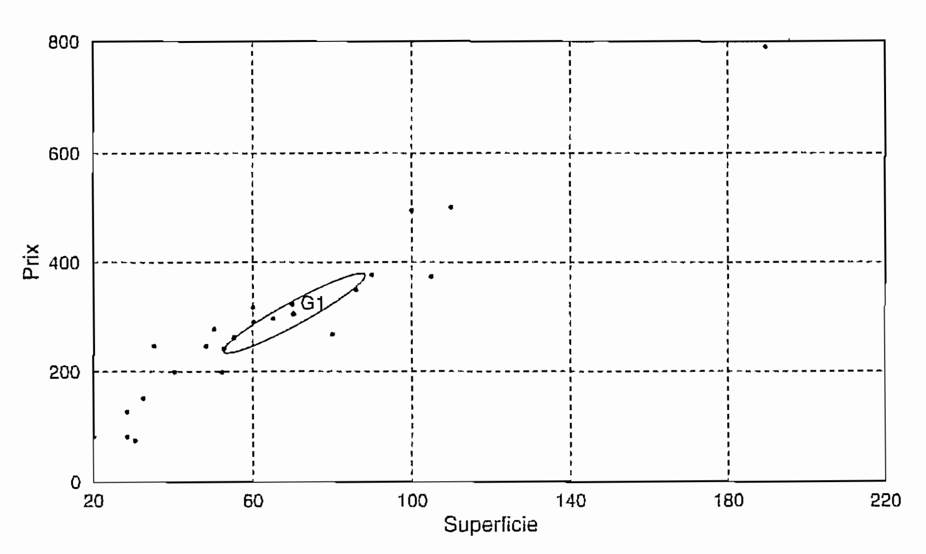


FIGURE 13.5 Ellipse de confiance à 95 %

13.6 INTERVALLES DE PRÉDICTION ET DE TOLÉRANCE

13.6.1 Prévision d'une valeur d'une loi normale

Lorsque m et σ sont connus, l'intervalle de probabilité 0.95 à risques symètriques pour une valeur isolée est $m \pm 1.96\sigma$. Supposons maintenant que m soit inconnu et estimé par la moyenne d'un n-échantillon $\bar{x}$. Soit X une nouvelle observation, ***indépendante*** des n précédentes. L'écart entre X et $\bar{X}$ suit alors une loi normale :

$$N\left(0\,;\sigma\sqrt{1+\frac{1}{n}}\right) \text{ car } V(X-\bar{X}) = \sigma^2 + \frac{\sigma^2}{n}.$$

On en déduit l'intervalle de prévision pour une valeur future $\bar{x} \pm 1.96\sigma\sqrt{1+\dfrac{1}{n}}$

Lorsque σ est aussi inconnu, en appliquant la méthode de Student, on trouve aisément que $\dfrac{X-\bar{X}}{S\sqrt{\dfrac{n+1}{n-1}}} = \dfrac{X-\bar{X}}{S^*\sqrt{\dfrac{n+1}{n}}} = T_{n-1}$ d'où l'intervalle $\bar{x} \pm ts\sqrt{\dfrac{n+1}{n-1}}$.

L'intervalle précédent est une estimation de l'intervalle $m \pm 1.96\sigma$. On voit aisément qu'il est plus large et converge vers lui lorsque n augmente indéfiniment. On l'appelle intervalle de tolérance sans niveau de confiance, car il existe aussi des intervalles de tolérance avec niveau de confiance $1-\alpha$, tels que l'intervalle $m \pm 1.96\sigma$ soit contenu avec une probabilité $1-\alpha$ dans l'intervalle de tolérance. Les intervalles de tolérance avec

niveau de confiance sont plus larges que les intervalles de tolérance sans niveau de confiance. Les formules sont plus complexes et nous renvoyons le lecteur intéressé à l'ouvrage de Hahn & Meeker (1991).

13.6.2 Ellipsoïde de tolérance pour une distribution normale $N_p(\mu ; \Sigma)$

De manière similaire, l'ellipsoïde d'équation $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = k$, où k est le fractile de niveau $1 - \alpha$ d'un χ^2_p, est un domaine de probabilité $1 - \alpha$ pour $\mathbf{x}$.

Si $\boldsymbol{\mu}$ est estimé par $\mathbf{g}$, centre de gravité d'un nuage de n réalisations indépendantes de X, alors $\mathbf{x} - \mathbf{g}$ suit une loi $N_p\left(0 ; \boldsymbol{\Sigma}\left(1 + \frac{1}{n}\right)\right)$. Si l'on estime de plus $\boldsymbol{\Sigma}$ par la la matrice de variance du nuage V, en appliquant les formules du chapitre précédent, on trouve que :

$$\boxed{(\mathbf{x} - \mathbf{g})'\mathbf{V}^{-1} (\mathbf{x} - \mathbf{g}) = \frac{(n-1)p}{n-p} \frac{n+1}{n} F(p ; n-p)}$$

ce qui donne l'équation de l'ellipsoïde de tolérance en remplaçant la variable de Fisher par son fractile.

En reprenant les données des 24 appartements parisiens, on trouve l'ellipse de tolérance suivante avec $p = 2$. On distingue clairement l'existence de deux points atypiques.

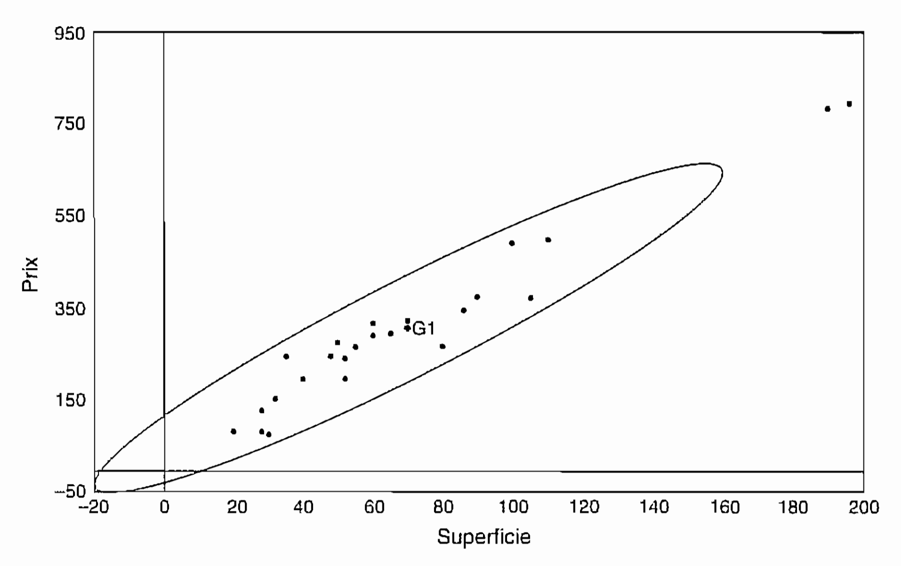


FIGURE 13.6 Ellipse de tolérance à 95 %

Les ellipses de tolérance sont très utiles en analyse discriminante.

13.7 ESTIMATION BAYÉSIENNE

Le point de vue bayésien ne fait pas de distinction de nature entre paramètres et observations : ce sont des variables aléatoires. Le probléme de l'estimation est alors résolu (en théorie du moins) de façon simple et élégante : il suffit de calculer la distribution *a posteriori* des paramétres sachant les observations.

13.7.1 Présentation

Soit un n-échantillon dé variables indépendantes et identiquement distribuées telles que leurs densités conditionnelles X_i/θ soient $f(x_i ; \theta)$.

Si l'on note comme d'habitude $L(\mathbf{x} ; \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$, la vraisemblance (ici conditionnelle), la loi conjointe des observations et du paramètre $(X_1, X_2, \ldots, X_n, \theta)$ est $L(\mathbf{x} ; \theta)g(\theta)$ où $g(\theta)$ est la densité *a priori* de θ.

La loi *a posteriori* du paramétre est $g(\theta/\mathbf{x}) = \dfrac{L(\mathbf{x} ; \theta)g(\theta)}{f(\mathbf{x})}$. Elle est donc proportionnelle au produit de la vraisemblance par la densité *a priori*.

On peut donc en déduire des régions probables pour θ, analogues bayésiens des régions de confiance classiques, mais aussi des estimations ponctuelles : il suffit de calculer un paramètre de tendance centrale de la loi *a posteriori*, le plus souvent l'espérance, mais aussi le mode ou la médiane.

13.7.2 Estimation bayésienne de la moyenne μ d'une loi normale de variance connue

On suppose ici que la loi de X/μ est une $N(\mu ; \sigma)$ et que la loi *a priori* de μ est une $N(\mu_0 ; \tau)$. Un calcul simple montre que la loi *a posteriori* de $\mu/X_1, X_2, \ldots, X_n$ est une loi normale d'espérance $E(\mu/\mathbf{x}) = \dfrac{\dfrac{\sigma^2}{n}\mu_0 + \tau^2\overline{X}}{\dfrac{\sigma^2}{n} + \tau^2}$ et de variance $V(\mu/\mathbf{x}) = \dfrac{\dfrac{\tau^2\sigma^2}{n}}{\dfrac{\sigma^2}{n} + \tau^2}$.

L'espérance *a posteriori* de μ est donc une moyenne pondérée de l'espérance *a priori* et de la moyenne empirique des observations. Si l'on introduit le concept de ***précision*** qui est l'inverse de la variance, la précision *a priori* est $\eta_1 = \dfrac{1}{\tau^2}$, la précision de la moyenne empirique est $\eta_2 = \dfrac{n}{\sigma^2}$.

On voit alors que $E(\mu/\mathbf{x}) = \dfrac{\eta_1\mu_0 + \eta_2\overline{X}}{\eta_1 + \eta_2}$ et $\dfrac{1}{V(\mu/\mathbf{x})} = \eta_1 + \eta_2$. La précision de l'estimateur bayésien est donc la somme de la précision de l'estimation *a priori* et de celle de la moyenne empirique, l'estimateur bayésien est alors la moyenne des deux estimations (*a priori* et empirique) pondérées par les précisions. Si l'information *a priori* sur le

paramètre est très précise, les observations ne la modifient guère. Si la précision *a priori* tend vers zéro, ou si n tend vers l'infini, on retrouve l'estimateur classique $\overline{X}$.

13.7.3 Estimation bayésienne d'une proportion p

Illustrons ce cas par un exemple issu du contrôle de qualité : on est amené à estimer la probabilité p qu'une marchandise soit défectueuse à partir de l'observation du nombre de marchandises défectueuses X dans un lot de n marchandises.

Pour une valeur donnée de p, X suit une loi binomiale $\mathcal{B}(n\,;p)$. L'ensemble des valeurs possibles de p peut être probabilisé si des expériences antérieures ont permis d'étudier les variations de p. Tout se passe donc comme si p était une réalisation d'une variable π à valeurs dans $[0\,;1]$ que l'on supposera de densité $g(p)$ (loi *a priori*).

On a donc le modèle suivant : la loi conditionnelle de $X/\pi = p$ est une $\mathcal{B}(n\,;p)$ et la loi marginale de π de densité $g(p)$. On cherche en général à déduire p de la valeur de X. Il faut donc pour cela trouver la loi de probabilité *a posteriori* de π ou loi de $\pi/X = x$.

La formule de Bayes donne :

$$f(p/X = x) = \frac{P(X = x/\pi = p)g(p)}{P(X = x)}$$

soit : $$f(p/X = x) = \frac{C_n^x p^x(1-p)^{n-x}\,g(p)}{\displaystyle\int_0^1 C_n^x p^x(1-p)^{n-x}\,g(p)\mathrm{d}p} = \frac{p^x(1-p)^{n-x}\,g(p)}{\displaystyle\int_0^1 p^x(1-p)^{n-x}\,g(p)\mathrm{d}p}$$

on pourra alors estimer p en choisissant la valeur la plus probable *a posteriori* ou la valeur moyenne *a posteriori*.

Si π suit une loi bêta de paramètre a et b on a :

$$g(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}$$

d'où : $$g(p/X = x) = \frac{p^{a+x-1}(1-p)^{n+b-x-1}}{\displaystyle\int_0^1 p^{a+x-1}(1-p)^{n+b-x-1}\mathrm{d}p}$$

donc la loi de $\pi/X = x$ est une loi bêta de paramètres $a + x$ et $n + b - x$.

L'espérance *a posteriori* vaut alors $\dfrac{a+x}{a+b+n}$.

Tout se passe donc comme si l'on avait effectué $a + b$ expériences suplémentaires ayant mené à a défectueux.

Le choix des paramètres a et b de la loi bêta se fait en général à partir de considérations sur la valeur la plus probable *a priori* et son incertitude.

Si l'on choisit $a = b = 1$ ce qui correspond à une distribution uniforme de π sur $[0\,;1]$ (toutes les valeurs de p sont *a priori* équiprobables) on trouve comme estimation de p

soit $\frac{x+1}{n+2}$ (espérance *a posteriori*) soit x/n (mode ou valeur de p correspondant au maximum de $g(p/x)$).

Remarquons que la loi marginale de X peut s'obtenir aisément :

$$P(X = x) = \int_0^1 C_n^x p^x (1-p)^{n-x}\, dp = C_n^x \int_0^1 p^x (1-p)^{n-x}\, dp$$

$$= C_n^x \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{n!}{x!(n-x)!} \frac{x!(n-x)!}{(n+1)!}$$

$$P(X = x) = \frac{1}{n+1}$$

La loi de X est alors la loi discrète uniforme sur $[0 ; 1 ; \ldots ; n]$.

13.7.4 Généralisation

Les deux cas précédents se résolvaient simplement car les lois *a priori* et conditionnelles permettaient de trouver la loi *a posteriori* dans la même famille que la loi *a priori* : les lois *a priori* et conditionnelles sont dites « conjuguées ». Cette facilité mathématique disparaît si l'on travaille avec des lois quelconques et les calculs des distributions *a posteriori* deviennent impossibles à mener analytiquement. Il faut alors recourir à des techniques spécifiques de simulation pour obtenir de bonnes approximations des lois *a posteriori* (C. Robert, 2001).

On peut considérer la statistique bayésienne comme un raffinement de la statistique paramétrique et lui appliquer les mêmes critiques sur le caractère plus ou moins réaliste des modèles. De plus le choix de la distribution *a priori* peut donner lieu à des divergences entre spécialistes et reste fatalement subjectif (voir la discussion sur la nature des probabilités au chapitre 1). Il n'en reste pas moins que cette approche est la seule qui permette d'incorporer de l'information préalable et se révèle très utile dans des cas limites comme des essais de fiabilité où on ne constate que très peu de défaillances (voire même aucune) sur n essais : les estimations classiques du taux de défaillance sont alors impossibles ou très imprécises.

13.8 NOTIONS SUR L'ESTIMATION ROBUSTE

La théorie classique de l'estimation permet de déterminer les estimateurs optimaux pour une famille de lois de probabilité définie à l'avance. Ces estimateurs dépendent en général fortement de la loi hypothétique : si celle-ci n'est pas correcte, les estimateurs ne le seront pas. On peut donc chercher des classes d'estimateurs relativement insensibles à des modifications des lois *a priori* : c'est un premier type de robustesse. Un deuxième type de robustesse concerne l'insensibilité à des valeurs « aberrantes » : la moyenne arithmétique est sans doute le meilleur estimateur de l'espérance pour une vaste classe de lois mais elle est très sensible aux grandes valeurs. L'attention des théoriciens et des praticiens a donc été attirée sur la recherche d'estimateurs robustes en particulier pour la valeur centrale d'une distribution.

On se préoccupera ici d'estimer la valeur centrale m d'une distribution symétrique.

La moyenne arithmétique $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est très sensible aux valeurs extrêmes : on obtiendra un estimateur robuste de m en éliminant les valeurs extrêmes. On appelle ainsi moyenne tronquée d'ordre α la moyenne arithmétique obtenue en éliminant les αn plus grandes et plus petites valeurs (une valeur généralement recommandée est $\alpha = 15\ \%$).

La médiane est le cas extrême d'une moyenne tronquée ($\alpha = 50\ \%$) et est très robuste.

Au lieu d'éliminer les αn plus grandes et plus petites valeurs, on peut les rendre toutes égales aux dernières valeurs prises en compte : c'est la "winsorization".

Une autre approche est celle des M-estimateurs introduits par P. Huber : on cherche ici μ qui minimise une fonction du type :

$$\sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{s}\right)$$

où s est un estimateur robuste de la dispersion ce qui revient à annuler $\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{s}\right)$ où $\psi = \rho'$.

On retrouve la moyenne arithmétique avec $\rho(x) = x^2$, la médiane avec $\rho(x) = |x|$.

Les estimateurs du maximum de vraisemblance sont des cas particuliers de M-estimateurs avec : $\rho(x) = -\ln f(x)$ et $\psi(x) = -\frac{f'(x)}{f(x)}$.

Remarquons que le M-estimateur μ peut s'écrire comme une moyenne pondérée des observations :

$$\mu = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

où w_i dépend des données $w_i = \frac{\psi(x_i - \mu)}{x_i - \mu}$.

Pour la moyenne arithmétique $\bar{x}$ $\psi(x) = x$.

Pour la médiane $\psi(x) = 1$ si $x > 0$ et $\psi(x) = -1$ si $x < 0$.

On obtiendra toute une famille d'estimateur en utilisant diverses formes de ψ :

$$\psi(x) = x\left(1 - \frac{x^2}{c^2}\right)^2 \quad \text{pour } |x| \leq c \qquad \text{(Tukey)}$$

$$\psi(x) = \begin{cases} -k & \text{si } x < -k \\ x & \text{si } |x| < k \\ +k & \text{si } x > k \end{cases} \quad \text{(Huber)}$$

Ces estimateurs sont obtenus par un processus de calcul itératif.

L'estimation robuste de dispersion s est prise généralement égale à la médiane des écarts absolus à la médiane.

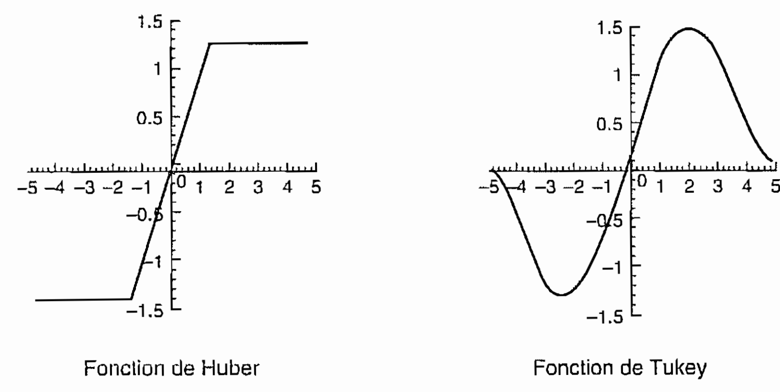


FIGURE 13.7

Dans l'exemple du chapitre 5, la variable « taux de taxe d'habitation » a une moyenne arithmétique de 17.7707, mais présentait quelques valeurs extrêmes.

Les estimations robustes sont :

Moyenne tronquée à 5 % :	17.6182
Estimateur de Huber avec $k = 1.339$:	17.8149
Estimateur de Tukey avec $c = 4.685$:	17.6872

13.9 ESTIMATION DE DENSITÉ

La densité $f(x)$ d'une variable continue donne une information visuelle importante sur la répartition des valeurs. Nous présentons ci-dessous les éléments de la théorie de l'estimation de la densité en l'absence de tout modèle paramétrique : on parle d'estimation fonctionnelle ou non-paramétrique. On supposera que $f(x)$ est une fonction continue.

La plupart des démonstrations seront omises au profit d'une présentation pratique. Le lecteur intéressé se reportera aux ouvrages de M. Delecroix et B. Silverman cités en bibliographie.

13.9.1 Généralités

Pour tout point x on cherche une fonction des observations $(x_1, x_2, \ldots)$ $\hat{f}_n(x)$ possédant les propriétés d'une estimation de la densité inconnue $f(x)$. Il semble légitime de souhaiter que :

- $\hat{f}_n(x)$ soit une densité (positive, d'intégrale égale à 1)
- $\hat{f}_n(x)$ soit convergent
- $\hat{f}_n(x)$ soit sans biais

Un résultat d'apparence paradoxale est que la propriété d'être sans biais est impossible à satisfaire : il n'existe pas d'estimateur sans biais en tout point x de la densité.

Pour la convergence, on se préoccupera non seulement de la convergence en tout point mais aussi de la convergence uniforme afin de borner l'erreur d'estimation maximale $\sup_x|\hat{f}_n(x) - f(x)|$.

L'erreur quadratique moyenne intégrée (MISE en anglais) est souvent utilisée pour mesurer l'écart quand n est fini entre l'estimateur et la densité inconnue :

$$\text{MISE} = E\left(\int_{\mathbb{R}}\left(\hat{f}_n(x) - f(x)\right)^2 \mathrm{d}x\right)$$

13.9.2 De l'histogramme à la fenêtre mobile

Considérons des histogrammes à classes d'égales amplitudes h. L'histogramme est l'estimateur de la densité le plus élémentaire.

Pour qu'il soit convergent, il faut faire tendre la largeur de classe vers 0 quand n tend vers l'infini, mais il ne faut pas que h tende vers zéro trop vite pour que l'effectif par classe puisse quand même tendre vers l'infini et assurer la convergence au point x. Il faut que $nh \to \infty$ ce qui peut être assuré par $h = \dfrac{a}{\sqrt{n}}$.

Mais en pratique n est fini et l'histogramme souffre de défauts évidents : il est discontinu, et constitue donc une approximation rustique d'une fonction continue. De plus par construction, tous les points d'un intervalle ont la même densité estimée, ce qui n'est pas réaliste.

Une première amélioration due à Rosenblatt est la méthode de la « fenêtre mobile » : on construit autour de chaque x une classe de longueur h centrée sur x : $[x - h/2 ; x + h/2]$ et on fait ensuite varier x. L'estimation en x est $\hat{f}_n(x) = \dfrac{n_x}{nh}$ où n_x est le nombre d'observations tombant dans la classe.

Cet estimateur reste cependant discontinu, car n_x varie de plus ou moins une unité à chaque fois que x correspond à une des valeurs x_i de l'échantillon.

L'exemple suivant montre le résultat pour les données déjà étudiées au chapitre 5 :

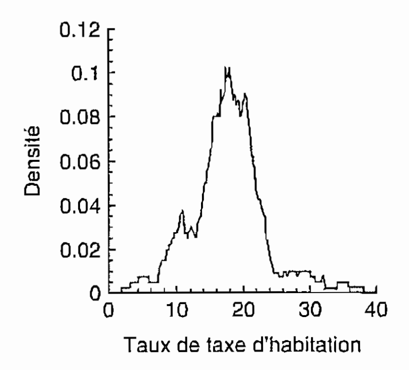


FIGURE 13.8

13.9.3 La méthode du noyau (Parzen)

Remarquons que l'estimateur de la fenêtre mobile peut s'écrire :

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

où le « noyau » K est la fonction indicatrice de l'intervalle $[-1/2 ; 1/2]$.

$$\begin{cases} K(u) = 1 \text{ si } -\dfrac{1}{2} \leq u \leq \dfrac{1}{2} \\ K(u) = 0 \text{ sinon} \end{cases}$$

$\hat{f}_n(x)$ est donc une moyenne arithmétique de fonctions donnant à chaque observation x_i un poids $1/h$ si elle appartient à l'intervalle centré sur x.

C'est parce que K est discontinue que $\hat{f}_n(x)$ l'est. Pour obtenir une estimation continue, on prendra une fonction noyau $K(u)$ continue ; on la choisira de plus paire par raison de symétrie, décroissante quand u s'éloigne de zéro. $\hat{f}_n(x)$ est alors une moyenne de fonctions donnant à chaque observation x_i un poids d'autant plus petit que $|x_i - x|$ est grand. Si K est une densité alors $\hat{f}_n(x)$ le sera également.

Les noyaux les plus couramment utilisés sont :

- le noyau triangulaire $K(u) = 1 - |u|$ si $-1 \leq u \leq 1$
- le noyau parabolique d'Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)$ si $-1 \leq u \leq 1$
- le biweight de Tukey $K(u) = \frac{15}{16}(1 - u^2)^2$ si $-1 \leq u \leq 1$
- le noyau sinusoïdal $K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)$ si $-1 \leq u \leq 1$
- le noyau gaussien $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$

Les noyaux à support borné nécessitent moins de calculs que le noyau gaussien. La constante de lissage h détermine la régularité de $\hat{f}_n(x)$. Comme pour la largeur des classes d'un histogramme un h trop grand lisse trop et un h trop petit conduit à une estimation très chaotique alors que le choix du noyau n'est pas crucial.

Ainsi pour les mêmes données que précédemment on trouve les estimations suivantes pour le noyau sinusoïdal avec trois largeurs de fenêtre 10 %, 20 % et 30 % de l'étendue de X) :

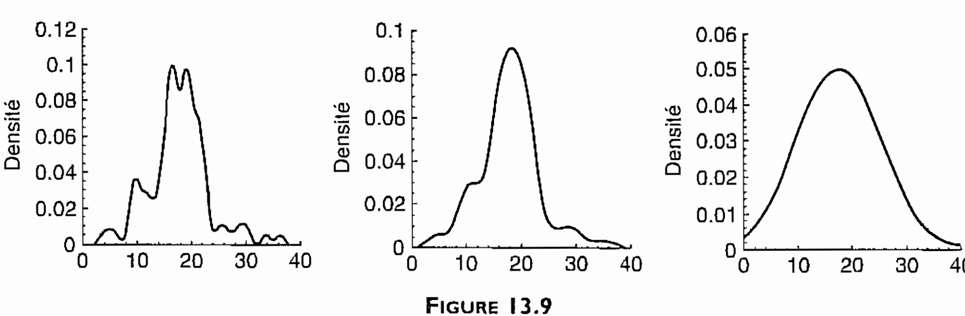


FIGURE 13.9

Le choix de la valeur « optimale » de h a fait l'objet d'une abondante littérature. Outre le choix subjectif par inspection visuelle (les logiciels permettent souvent de faire varier h en continu avec un curseur), mentionnons :

- la règle gaussienne $h = \hat{\sigma}[4/(3n)]^{1/5}$
- la règle de Silverman $h = \dfrac{0.9\min[\hat{\sigma}\,;(Q_3 - Q_1)/1.34]}{n^{1/5}}$
- la validation croisée généralisée où on cherche en quelque sorte une estimation du maximum de vraisemblance mais en enlevant x_i pour l'estimation en x_i $\max\limits_h \prod\limits_{i=1}^{n} \hat{f}_{n-1}^{-i}(x_i)$.

14 Les tests statistiques

14.1 INTRODUCTION

14.1.1 Les faiseurs de pluie

Des relevés effectués pendant de nombreuses années ont permis d'établir que le niveau naturel des pluies dans la Beauce en millimètres par an suit une loi de Laplace-Gauss LG(600, 100).

Des entrepreneurs, surnommés faiseurs de pluie, prétendaient pouvoir augmenter de 50 mm le niveau moyen de pluie, ceci par insémination des nuages au moyen d'iodure d'argent. Leur procédé fut mis à l'essai entre 1951 et 1959 et on releva les hauteurs de pluies suivantes :

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	510	614	780	512	501	534	603	788	650

Que pouvait-on en conclure ? Deux hypothèses s'affrontaient : ou bien l'insémination était sans effet, ou bien elle augmentait réellement le niveau moyen de pluie de 50 mm.

Ces hypothèses pouvaient se formaliser comme suit, si m désigne l'espérance mathématique de X variable aléatoire égale au niveau annuel de pluie :

$$\begin{cases} H_0 : m = 600 \text{ mm} \\ H_1 : m = 650 \text{ mm} \end{cases}$$

Les agriculteurs hésisant à opter pour le procédé forcément onéreux des faiseurs de pluie tenaient pour l'hypothèse H_0 et il fallait donc que l'expérience puisse les convaincre ; c'est-à-dire que les faits observés contredisent nettement la validité de l'hypothèse H_0 dite « hypothèse nulle » (H_1 s'appelle l'hypothèse alternative). Les agriculteurs n'étaient donc décidés à abandonner H_0 qu'en présence de faits expérimentaux traduisant une éventualité improbable compte tenu de H_0.

Ils choisirent $\alpha = 0.05$ comme niveau de probabilité, c'est-à-dire qu'ils étaient prêts à accepter H_1 si le résultat obtenu faisait partie d'une éventualité improbable qui n'avait que 5 chances sur 100 de se produire. Autrement dit, ils admettaient implicitement que

des événements rares ne sauraient se produire sans remettre en cause le bien-fondé de l'hypothèse de départ H_0 ; ce faisant, ils assumaient le risque de se tromper dans 5 cas sur 100, cas où précisément les événements « rares » arrivent quand même.

Comment décider ? Puisqu'il s'agit de « tester » la valeur m il est naturel de s'intéresser à $\overline{X}$ moyenne des observations qui nous apporte le plus de renseignements sur m. $\overline{X}$ est la « variable de décision ».

Si H_0 est vraie, comme l'expérience a porté sur $n = 9$ ans, $\overline{X}$ doit suivre une loi de Laplace-Gauss $LG\left(600, \dfrac{100}{\sqrt{9}}\right)$.

En principe, de grandes valeurs de $\overline{X}$ sont improbables et on prendra comme règle de décision la suivante :

Si $\overline{X}$ est trop grand, c'est-à-dire si $\overline{X}$ est supérieur à un seuil k qui n'a que 5 chances sur 100 d'être dépassé, on optera pour H_1 avec une probabilité 0.05 de se tromper.

Si $\overline{X} < k$ on conservera H_0 faute de preuves suffisantes. Il est facile de calculer k grâce aux tables et on trouve :

$$k = 600 + \frac{100}{3} 1.64 = 655 \quad \text{(fig. 14.1)}$$

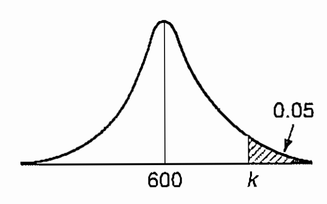


Figure 14.1

La règle de décision est donc la suivante :

- si $\overline{X} > 655$ mm, repousser H_0 et accepter H_1 ;
- si $\overline{X} < 655$ mm, conserver H_0.

L'ensemble d'événements $\{\overline{X} > 655\}$ s'appelle la ***région critique*** ou région de rejet de H_0. L'ensemble complémentaire $\{\overline{X} < 655\}$ s'appelle la région d'acceptation de H_0.

Or, les données relevées indiquent que $\overline{x} = 610.2$ mm. La conclusion était donc de conserver H_0 ; c'est-à-dire que l'insémination était sans effet notable sur le niveau des pluies : les valeurs observées pouvaient donc être dues au hasard en l'absence de toute influence de l'iodure d'argent.

Cependant, rien ne dit que conserver H_0 mette à l'abri de se tromper : en effet, les faiseurs de pluie ont peut-être raison, mais on ne s'en est pas aperçu.

Il y avait deux manières de se tromper : croire les faiseurs de pluie, alors qu'ils n'étaient pour rien dans le résultat obtenu (probabilité $\alpha = 0.05$) ; ne pas croire les faiseurs de pluie, alors que leur méthode est bonne et que seul le hasard (malencontreux pour eux), dû au faible nombre d'observations, a donné des résultats insuffisants pour convaincre les agriculteurs.

Supposons que les faiseurs de pluie ont raison, alors $\overline{X} \in LG\left(650, \frac{100}{3}\right)$. On commet une erreur chaque fois que $\overline{X}$ prend une valeur inférieure à 655 mm, c'est-à-dire avec une probabilité :

$$\beta = P\left(U < \frac{655 - 650}{100/3}\right) = P(U < 0.15) \qquad \beta = 0.56$$

ce qui est considérable.

α s'appelle le risque de première espèce ; β s'appelle le risque de deuxième espèce.

On aura remarqué au cours de cet exemple le rôle particulier joué par H_0 : si la forme de la région critique $\overline{X} > k$ est indiquée par la nature de H_1 (650 plus grand que 600) la valeur de k ne dépend que de H_0.

Les deux hypothèses ne jouent pas des rôles symétriques, k est déterminé par H_0 et α ; β est déterminé par la considération supplémentaire de H_1.

14.1.2 Les grandes catégories de tests

On peut classer les tests selon leur objet (ajustement, indépendance, de moyenne, de variance, etc.), ainsi qu'il est fait dans la suite du livre ou selon leurs propriétés mathématiques : on parle ainsi de tests paramétriques ou non, de tests robustes, de tests libres.

Un test est dit ***paramétrique*** si son objet est de tester certaine hypothèse relative à un ou plusieurs paramètres d'une variable aléatoire de loi spécifiée ou non : le paragraphe 14.2 en donne des exemples. Dans la plupart des cas, ces tests sont basés sur la loi normale et supposent donc explicitement l'existence d'une variable aléatoire de référence X suivant une loi LG. La question se pose alors de savoir si les résultats restent encore valables lorsque X n'est pas normale : si les résultats sont valables on dit que le test en question est ***robuste***. La robustesse d'un test par rapport à un certain modèle est donc la qualité de rester relativement insensible à certaines modifications du modèle : on constatera que les tests de moyenne ou de non corrélation sont robustes.

Une catégorie particulièrement intéressante de tests robustes est la classe des ***tests libres*** (en anglais ***distribution free***) : il s'agit de tests valables quelle que soit la loi de la variable aléatoire étudiée, donc valables en particulier lorsque l'on ignore tout de cette loi (cas très fréquent en pratique) ; on peut dire qu'il s'agit de tests robustes par rapport à la loi de probabilité. Exemple : les tests d'ajustement du χ^2. Ces tests sont bien souvent des tests ***non paramétriques*** mais pas nécessairement (tests de moyenne).

Pour les tests paramétriques on distingue généralement hypothèses simples et hypothèses composites :

- ***une hypothèse simple*** est du type $H : \theta = \theta_0$ où θ_0 est une valeur isolée du paramètre ;
- ***une hypothèse composite*** est du type $H : \theta \in A$ où A est une partie de $\mathbb{R}$ non réduite à un élément.

La plupart des hypothèses composites se ramènent aux cas : $\theta > \theta_0$ ou $\theta < \theta_0$ ou $\theta \neq \theta_0$.

En fait, on construira les régions critiques en utilisant la valeur θ_0 seule. Lorsque l'hypothèse alternative est composite, la puissance du test est variable et on parle de fonction puissance $1 - \beta(\theta)$.

14.2 THÉORIE CLASSIQUE DES TESTS

Un test est un mécanisme qui permet de trancher entre deux hypothèses au vu des résultats d'un échantillon.

Soient H_0 et H_1 ces deux hypothèses, dont une et une seule est vraie. La décision aboutira à choisir H_0 ou H_1. Il y a donc 4 cas possibles schématisés dans le tableau 14.1 avec les probabilités correspondantes :

TABLEAU 14.1

Décision \ Vérité	H_0	H_1
H_0	$1 - \alpha$	β
H_1	α	$1 - \beta$

14.2.1 Risques et probabilités d'erreur

α et β sont les probabilités d'erreur de première et deuxième espèce :

- α probabilité de choisir H_1 alors que H_0 est vraie ;
- β probabilité de conserver H_0 alors que H_1 est vraie.

Ces erreurs correspondent à des risques différents en pratique ; ainsi dans l'exemple des faiseurs de pluie le risque de première espèce consiste à acheter un procédé d'insémination inefficace ; le risque de deuxième espèce à laisser perdre une occasion d'augmenter le niveau de pluie et peut-être de récoltes plus abondantes.

Dans la pratique des tests statistiques, il est de règle de se fixer α comme donné (les valeurs courantes sont par exemple 0.05, 0.01 ou 0.1) de préférence en fonction du risque de première espèce couru, ce qui fait jouer à H_0 un rôle prééminent.

Le choix de H_0 est dicté par des motifs assez variables :

- puisqu'on ne veut pas abandonner trop souvent H_0, H_0 doit être une hypothèse solidement établie et qui n'a pas été contredite jusqu'à présent par l'expérience ;
- H_0 est une hypothèse à laquelle on tient particulièrement pour des raisons qui peuvent être subjectives ;
- H_0 correspond à une hypothèse de prudence ; exemple : test de l'innocuité d'un vaccin ; il est prudent de partir d'une hypothèse défavorable au nouveau produit ;
- H_0 est la seule hypothèse facile à formuler ; exemple : tester $m = m_0$ contre $m \neq m_0$; il est évident que seule $H_0 : m = m_0$ permettra d'effectuer des calculs.

α étant fixé, β sera déterminé comme résultat d'un calcul (ceci n'est possible que si l'on connaît les lois de probabilités sous H_1).

Cependant il faut savoir que β varie en sens contraire de α. Si l'on veut diminuer α risque d'erreur de première espèce, on augmente $1 - \alpha$ probabilité d'accepter H_0, si H_0 est vraie ; mais surtout on est conduit à une règle de décision plus stricte qui aboutit à n'abandonner H_0 que dans des cas rarissimes donc à conserver H_0 bien souvent à tort.

A force de ne pas vouloir abandonner H_0 on finit par la garder presque tout le temps, donc on augmente β.

$1 - \beta$ est la probabilité d'opter pour H_1 en ayant raison. ***$1 - \beta$ s'appelle « puissance du test ».***

α étant fixé, il importe de choisir une variable de décision : variable qui doit apporter le maximum d'informations sur le problème posé et dont la loi sera différente selon que H_0 ou H_1 est vraie (sinon elle ne servirait à rien). Il faut que sa loi soit entièrement connue au moins si H_0 est vraie.

La région critique W est l'ensemble des valeurs de la variable de décision qui conduisent à écarter H_0 au profit de H_1. La forme de la région critique est déterminée par la nature de H_1, sa détermination exacte se fait en écrivant que :

$$P(W|H_0) = \alpha$$

La région d'acceptation est son complémentaire $\overline{W}$ et l'on a donc :

$$P(\overline{W}|H_0) = 1 - \alpha \qquad \text{et} \qquad P(W|H_1) = 1 - \beta$$

La construction d'un test n'est rien d'autre que la détermination de la région critique, cette détermination se faisant sans connaître le résultat de l'expérience, donc *a priori*.

La démarche d'un test est la suivante (pour résumer) :

1) Choix de H_0 et H_1.
2) Détermination de la variable de décision.
3) Allure de la région critique en fonction de H_1.
4) Calcul de la région critique en fonction de α.
5) Calcul éventuel de la puissance $1 - \beta$.
6) Calcul de la valeur expérimentale de la variable de décision.
7) Conclusion : rejet ou acceptation de H_0.

14.2.2 Choix de la variable de décision et de la région critique optimales : la méthode de Neyman et Pearson

La façon de choisir la variable de décision n'a pas encore été élucidée dans les paragraphes précédents où nous nous sommes contentés de formaliser l'intuition.

Le choix de la meilleure variable de décision a été résolu théoriquement par les statisticiens J.Neyman et E.S.Pearson dans une série d'articles célèbres parus de 1933 à 1938.

Mais que veut dire « meilleure variable », ou plutôt ainsi que nous l'utiliserons désormais, région critique optimale ?

Nous cherchons la région critique optimale c'est-à-dire un domaine de $\mathbb{R}^n$ parmi l'ensemble de toutes les réalisations possibles de l'échantillon $(X_1, X_2, \ldots, X_n)$ dont la forme définira ensuite une variable statistique.

Il s'agit de maximiser la puissance $1 - \beta$ ceci pour une valeur donnée de α risque de première espèce.

Nous allons tout d'abord envisager le test entre deux hypothèses paramétriques simples puis nous généraliserons à d'autres types d'hypothèses.

Soit X une variable aléatoire de densité $f(x, \theta)$ où θ est un paramètre réel inconnu ; $L(\mathbf{x}, \theta)$ désignera en abrégé la densité de l'échantillon.

Il s'agit de tester : $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$

Supposons α connu. Soit W une région de $\mathbb{R}^n$ telle que :

$$\int_W L(\mathbf{x}, \theta_0)\mathrm{d}\mathbf{x} = \alpha = P(W|H_0)$$

Il s'agit de maximiser : $1 - \beta = \int_W L(\mathbf{x}, \theta_1)\mathrm{d}\mathbf{x} = P(W|H_1)$

Nous pouvons écrire : $1 - \beta = \int_W \frac{L(\mathbf{x}, \theta_1)}{L(\mathbf{x}, \theta_0)} L(\mathbf{x}, \theta_0)\mathrm{d}\mathbf{x}$

Théorème de Neyman et Pearson

La région critique optimale est définie par l'ensemble des points de $\mathbb{R}^n$ tels que :

$$\boxed{\frac{L(\mathbf{x} ; \theta_1)}{L(\mathbf{x} ; \theta_0)} > k_\alpha}$$

Démonstration

- S'il existe une constante k_α, telle que l'ensemble W des points de $\mathbb{R}^n$ où :

$$\frac{L(\mathbf{x} ; \theta_1)}{L(\mathbf{x} ; \theta_0)} > k_\alpha$$

soit de probabilité α sous H_0 : $P(W|H_0) = \alpha$, alors cette région W réalise le maximum de $1 - \beta$.

En effet soit W' une autre région de $\mathbb{R}^n$ telle que $P(W'|H_0) = \alpha$; W' diffère alors de W par des points où $\frac{L(\mathbf{x} ; \theta_1)}{L(\mathbf{x} ; \theta_0)} \leq k_\alpha$ (fig. 14.2). L'intégrale :

$$\int_W \frac{L(\mathbf{x} ; \theta_1)}{L(\mathbf{x} ; \theta_0)} L(\mathbf{x} ; \theta_0)\mathrm{d}\mathbf{x}$$

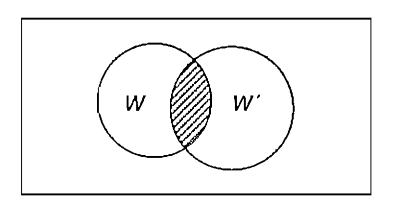


Figure 14.2

diffère de l'intégrale : $$\int_{W'} \frac{L(\mathbf{x}\,;\theta_1)}{L(\mathbf{x}\,;\theta_0)} L(\mathbf{x}\,;\theta_0)d\mathbf{x}$$

pour les parties non communes à W et W'.

W et W' ayant sous H_0 même mesure α, l'ensemble $(W - W')$ des points de W n'appartenant pas à W' a même mesure sous H_0 que l'ensemble $(W' - W)$ des points de W' n'appartenant pas à W.

L'intégrale : $$\int_{W-W'} \frac{L(\mathbf{x}\,;\theta_1)}{L(\mathbf{x}\,;\theta_0)} L(\mathbf{x}\,;\theta_0)d\mathbf{x}$$

est alors strictement supérieure à :

$$\int_{W'-W} \frac{L(\mathbf{x}\,;\theta_1)}{L(\mathbf{x}\,;\theta_0)} L(\mathbf{x}\,;\theta_0)d\mathbf{x}$$

comme intégrale, prise par rapport à la mesure $L(\mathbf{x}\,;\theta_0)d\mathbf{x}$ ***sur un ensemble de même mesure***, d'une fonction strictement supérieure ; le théorème de la moyenne indique en effet :

$$\int_{W'-W} \frac{L(\mathbf{x}\,;\theta_1)}{L(\mathbf{x}\,;\theta_0)} L(\mathbf{x}\,;\theta_0)d\mathbf{x} = \frac{L(\xi'\,;\theta_1)}{L(\xi'\,;\theta_0)} P(W' - W|H_0) \qquad \text{avec } \xi' \in W' - W$$

$$\int_{W-W'} \frac{L(\mathbf{x}\,;\theta_1)}{L(\mathbf{x}\,;\theta_0)} L(\mathbf{x}\,;\theta_0)d\mathbf{x} = \frac{L(\xi\,;\theta_1)}{L(\xi\,;\theta_0)} P(W - W'|H_0) \qquad \text{avec } \xi \in W - W'$$

ce qui démontre le point *a*) car :

$$\frac{L(\xi'\,;\theta_1)}{L(\xi'\,;\theta_0)} \leq k_\alpha < \frac{L(\xi\,;\theta_1)}{L(\xi\,;\theta_0)}$$

- Montrons que cette constante k_α existe.

Soit $A(K)$ la région de $\mathbb{R}^n$ où $L(\mathbf{x}, \theta_1) > KL(\mathbf{x}, \theta_0)$ et considérons $P(A(K)|H_0)$ qui est une fonction continue monotone de K, si X est à densité continue. Comme $L(\mathbf{x}, \theta_1)$ est toujours positif, car c'est une densité, on a $P(A(0)|H_0) = 1$. D'autre part si $K \to \infty$, avec une densité bornée on a $P(A(K)|H_0) \to 0$. Il existe donc une valeur intermédiaire k_α telle que $P(A(k_\alpha)) = \alpha$. ■

14.2.3 Étude de $1 - \beta$: puissance du test

Nous allons montrer que $1 - \beta > \alpha$. Un tel test est dit sans biais :

$$P(W|H_1) > P(W|H_0)$$

puisque : $$L(\mathbf{x}, \theta_1) > k_\alpha L(\mathbf{x}, \theta_0)$$

d'où : $$\int_W L(\mathbf{x}, \theta_1)d\mathbf{x} > k_\alpha \int_W L(\mathbf{x}, \theta_0)d\mathbf{x}$$

Si k_α est > 1 la proposition est triviale ; si k_α est < 1 nous allons montrer, ce qui est équivalent, que $\beta < 1 - \alpha$:

$$\beta = P(\overline{W}|H_1) \qquad \text{et} \qquad 1 - \alpha = P(\overline{W}|H_0)$$

$\overline{W}$ est tel que $\dfrac{L(\mathbf{x}, \theta_1)}{L(\mathbf{x}, \theta_0)} < k_\alpha$, donc :

$$\int_{\overline{W}} L(\mathbf{x}, \theta_1)\mathrm{d}\mathbf{x} < k_\alpha \int_{\overline{W}} L(\mathbf{x}, \theta_0)\mathrm{d}\mathbf{x}$$

ce qui démontre la proposition.

Convergence du test : On peut démontrer que si $n \rightarrow \infty$, $1 - \beta \rightarrow 1$.

Remarque : Comme $P(A(K))$ est une fonction monotone de K, on voit que si α diminue, k_α augmente ; donc diminuer le risque de première espèce α fait augmenter le risque de deuxième espèce $\beta 1 - \beta = P\left(\dfrac{L(\mathbf{x}, \theta_1)}{L(\mathbf{x}, \theta_0)} > k_\alpha \middle| H_1\right)$ est une fonction décroissante de k.

14.2.4 Tests et statistiques exhaustives

La considération d'une statistique exhaustive simplifie considérablement la pratique du test car alors la région critique en dépend exclusivement.

S'il existe une statistique exhaustive T pour θ, de densité $g(t, \theta)$, on a :

$$L(\mathbf{x}, \theta) = g(t, \theta)h(\mathbf{x})$$

Le test de Neyman et Pearson se réduit alors à :

$$\frac{g(t, \theta_1)}{g(t, \theta_0)} > k_\alpha$$

14.2.5 Exemple

Test de la moyenne d'une loi de Laplace-Gauss, d'écart-type σ connu :

$$H_0 : LG(m_0, \sigma) \qquad \text{contre} \qquad H_1 : LG(m_1, \sigma)$$

La statistique exhaustive pour m est $\bar{x}$ et :

$$g(\bar{x}, m) = \frac{1}{\sigma\sqrt{\dfrac{2\pi}{n}}} \exp\left(-\frac{1}{2}\left(\frac{\bar{x} - m}{\sigma/\sqrt{n}}\right)^2\right)$$

Le rapport des densités $\dfrac{g(\bar{x}, m_1)}{g(\bar{x}, m_0)}$ donne :

$$\frac{g(\bar{x}, m_1)}{g(\bar{x}, m_0)} = \exp\left(-\frac{n}{2\sigma^2}[(\bar{x} - m_1)^2 - (\bar{x} - m_0)^2]\right)$$

Écrire que $\dfrac{g(\bar{x}, m_1)}{g(\bar{x}, m_0)} > k_\alpha$ est équivalent à écrire que $(\bar{x} - m_0)^2 - (\bar{x} - m_1)^2 > k'_\alpha$ soit :

$$(m_1 - m_0)(2\bar{x} - m_0 - m_1) > k'_\alpha$$

Si $m_1 > m_0$ il est équivalent d'écrire $\bar{x} > k''_\alpha$.

Si $m_1 < m_0$ il est équivalent d'écrire $\bar{x} < k'''_\alpha$.

Ce résultat évident à l'intuition exprime que si $m_1 > m_0$, on rejettera H_0 si $\overline{X}$ est trop grand. On trouve la constante k en écrivant $P(\overline{X} > k \mid H_0) = \alpha$.

En représentant sur un même graphique les densités de $\overline{X}$ dans H_0 et dans H_1 on a la figure 14.3.

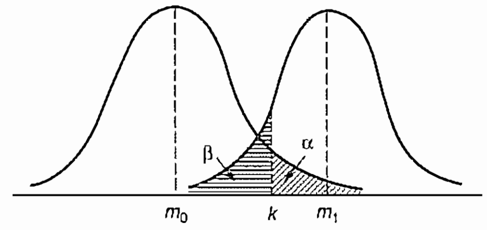


FIGURE 14.3

14.2.6 Tests entre hypothèses composites

14.2.6.1 Test d'une hypothèse simple contre une hypothèse composite

Exemples : 1) $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$ 2) $\begin{cases} \theta = \theta_0 \\ \theta \neq \theta_0 \end{cases}$

- La fonction puissance

L'hypothèse H_1 étant composée d'un ensemble de valeurs de θ, pour chaque valeur particulière de θ on peut calculer $1 - \beta(\theta)$, d'où une fonction, dite fonction puissance, décrivant les variations de $1 - \beta$ selon les valeurs de θ dans H_1.

La figure 14.4 montre la fonction puissance du test $H_0 : m = 600$ contre $H_1 : m > 600$ correspondant à l'exemple introductif.

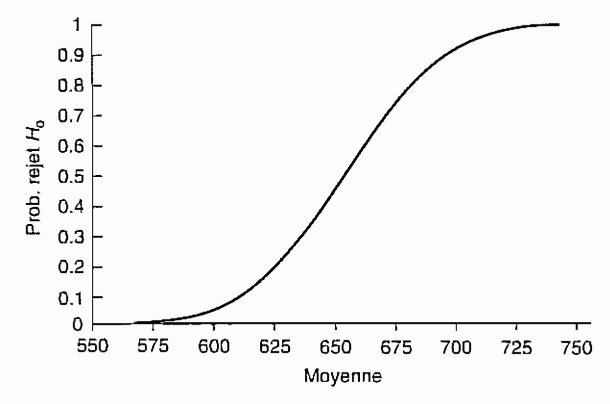


FIGURE 14.4 Fonction puissance. Test unilatéral $H_0 = 600$.

Note : la courbe donnant β en fonction du paramètre est appelée courbe d'efficacité.

- Tests UPP

Un test est dit uniformément le plus puissant (UPP) si, quelle que soit la valeur de θ appartenant à l'hypothèse alternative, sa puissance $1 - \beta(\theta)$ est supérieure à la puissance de tout autre test.

Exemple : Dans le test $H_0 : m = m_0$ contre $H_1 : m = m_1 > m_0$, on a pu remarquer que la région critique ne dépend pas explicitement de m_1 et donc que cette région critique est la même pour n'importe quel $m_1 > m_0$. Le test précédent est donc UPP pour $H_0 : m = m_0$ contre $H_1 : m > m_0$.

Il est évident cependant qu'il n'existe pas de test UPP pour $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ car, s'il en existait un il devrait être UPP pour les deux sous-hypothèses $H_1' : m > m_0$ et $H_1'' : m < m_0$. Or les tests de H_0 contre H_1' et H_0 contre H_1'' sont précisément UPP et différents l'un de l'autre.

14.2.6.2 Test entre deux hypothèses composites

Si H_0 est elle-même composite, α dépend de θ selon les valeurs de $\theta \in H_0$, et l'on devra exiger $\alpha(\theta) \leq \alpha$ donné.

L'existence de tests UPP pour les cas suivants :

$$\begin{cases} H_0 : \theta < \theta_0 \\ H_1 : \theta \geq \theta_0 \end{cases} \quad \text{et} \quad \begin{cases} H_0 : \theta \leq \theta_1 \text{ ou } \theta \geq \theta_2 \\ H_1 : \theta_1 < \theta \leq \theta_2 \end{cases}$$

est assurée par le théorème de Lehmann que nous ne démontrerons pas.

Ce théorème suppose l'existence d'une statistique G telle que le rapport $\dfrac{L(\mathbf{x} ; \theta_1)}{L(\mathbf{x} ; \theta_2)}$ est une fonction monotone croissante de G si $\theta_1 > \theta_2$ (théorème dit « du rapport de vraisemblance monotone »).

De telles statistiques sont fournies par les statistiques exhaustives des lois de forme exponentielle.

D'autre part il n'existe pas de tests UPP pour les cas $H_0 : \theta_1 \leq \theta \leq \theta_2$ contre $H_1 : \theta > \theta_2$ ou $\theta < \theta_1$ et *a fortiori* : $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$.

Dans les cas où il n'existe pas de tests UPP, on cherchera s'il existe de bons tests parmi une classe plus restreinte, celle des tests sans biais par exemple. Ainsi pour le test précédent il existe un test UPP sans biais s'il existe une statistique $G(\mathbf{x})$ répondant à la condition de Lehmann et la région critique est :

$$G(\mathbf{x}) < c_1 \qquad \text{ou} \qquad G(\mathbf{x}) > c_2$$

14.2.6.3 Test du rapport des vraisemblances maximales

Ce test est fort utile là où les méthodes précédentes ont échoué :

- Test de H_0

$\theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ où θ peut être un paramètre vectoriel de dimension p.

Posons :
$$\lambda = \frac{L(\mathbf{x}, \theta_0)}{\sup\limits_{\theta} L(\mathbf{x}, \theta)}$$

on a donc $0 \leq \lambda \leq 1$.

λ est intuitivement une statistique convenable pour un test car plus λ est grand, plus l'hypothèse H_0 est vraisemblable (principe du maximum de vraisemblance). Cela revient à remplacer θ par son estimation $\hat{\theta}$ par la méthode du maximum de vraisemblance.

La région critique du test sera : $\lambda < K$

THÉORÈME 1

La distribution de $-2 \ln \lambda$ est asymptotiquement celle d'un χ^2_p dans l'hypothèse H_0.

Démonstration : Nous la ferons pour $p = 1$. On a, en développant en série de Taylor :

$$\ln L(\mathbf{x}, \theta_0) - \ln L(\mathbf{x}, \hat{\theta}) = (\theta_0 - \hat{\theta})\frac{\partial}{\partial\theta} \ln L(\mathbf{x}, \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})\frac{\partial}{\partial\theta^2} \ln L(\mathbf{x}, \theta^*)$$

ou $\theta^* \in [\theta_0, \hat{\theta}]$.

Comme θ est l'estimateur du MV on a $\frac{\partial}{\partial\theta} \ln L(x, \hat{\theta}) = 0$, d'où :

$$-2 \ln \lambda = -(\theta_0 - \hat{\theta})^2 \frac{\partial}{\partial\theta^2} \ln L(\mathbf{x}, \theta^*)$$

Dans l'hypothèse $H_0 : \theta = \theta_0$, on sait que l'estimation du MV converge presque sûrement vers θ_0 ; donc $\theta^* \longrightarrow \theta_0$ et lorsque $n \to \infty$:

$$\frac{\partial^2 \ln L(x ; \theta^*)}{\partial\theta^2} \sim \frac{\partial^2 \ln f(x_i ; \theta)}{\partial\theta^2} = n\frac{1}{n}\sum \frac{\partial^2 \ln f(x_i ; \theta)}{\partial\theta^2}$$

Lorsque $n \to \infty$, la loi des grands nombres nous indique que :

$$\frac{1}{n}\sum \frac{\partial^2 \ln f(x_i ; \theta)}{\partial\theta^2} \to E\left[\frac{\partial^2 \ln f}{\partial\theta^2}\right] = I_1(\theta)$$

Donc : $$\frac{\partial^2 \ln L(x ; \theta^*)}{\partial\theta} \to nI_1(\theta) = I_n(\theta)$$

On a alors : $$-2 \ln \lambda \sim (\theta_0 - \hat{\theta})^2 I_n(\theta_0)$$

D'autre part on sait que $\frac{\theta_0 - \hat{\theta}}{\sqrt{1/I_n(\theta)}} \to LG(0 ; 1)$. Donc $(\theta_0 - \hat{\theta})^2 I_n(\theta_0) \to \chi^2_1$.

Note : Le domaine de définition doit être indépendant du paramètre.

THÉORÈME 2

Si $n \to \infty$, la suite des tests est convergente, c'est-à-dire que la puissance $1 - \beta \to 1$.

- Test entre deux hypothèses composites.

On formera $\lambda = \frac{\sup_{\theta \in H_0} L(\mathbf{x}, \theta)}{\sup_{\theta \in H_1} L(\mathbf{x}, \theta)}$ et on obtient les mêmes propriétés que précédemment.

14.2.7 Niveau de signification, risques, vraisemblance et approche bayésienne

Dans la théorie classique que nous venons d'exposer, issue des travaux de J. Neyman et E.S. Pearson, un test se présente sous forme d'une règle de décision binaire formulée *a priori*, c'est-à-dire avant d'avoir fait les observations, et intangible dès que le risque α a été fixé. Les données recueillies ne peuvent modifier cette règle et ne sont utilisées qu' *a posteriori* : on est, ou on n'est pas, dans la région critique. Cette manière de procéder interdit toute manipulation des résultats et garantit l'objectivité de la décision : elle convient bien dans des problèmes de réglementation ou de normalisation.

Il y a cependant quelques inconvénients :

- Seule une partie de l'information est utilisée.
- On aboutit parfois à des conclusions paradoxales :

Ainsi si on rejette H_0 avec $\alpha = 5$ %, avec les mêmes observations on la rejettera *a fortiori* si l'on avait choisi $\alpha = 20$ %. Quel est alors le risque de la rejeter à tort : 5 ou 20 % ? Ce genre de questions perturbe à bon droit le praticien.

Une pratique courante, utilisée notamment dans les logiciels, consiste alors à calculer le ***niveau de signification***, appelé « *p*-value » en anglais : c'est la probabilité de dépassement de la valeur observée de la variable de décision sous H_0. Ainsi dans l'exemple des faiseurs de pluie, le niveau de signification est $P(\overline{X} > 610.2/H_0) = 0.38$.

Cela veut dire que pour tout $\alpha < 0.38$ on conserve H_0. Cette valeur élevée est donc en faveur de l'hypothèse nulle : les données la confortent. Inversement plus le niveau de signification est faible, plus les données sont en faveur de l'hypothèse alternative et du rejet de H_0. La démarche classique de Neyman-Pearson revient simplement à comparer le niveau de signification avec le risque α, mais on a ici une information plus précise.

On aimerait pouvoir dire que H_0 est plus « probable » que H_1. Il faut se garder de telles expressions, dénuées de sens dans un contexte non-bayésien. Par contre on peut parler de la vraisemblance de chacune des deux hypothèses (du moins quand elles sont simples) L_0 et L_1.

Il faut noter ici que le test de Neyman-Pearson ne consiste pas à se prononcer en faveur de l'hypothèse la plus vraisemblable puisque la constante k_α du rapport des vraisemblances n'est pas égale à 1 : il faut que H_1 soit k_α fois plus vraisemblable que H_0, et k_α est généralement plus grand que 1, ce qui traduit le fait que H_0 est privilégiée. Dans l'exemple des faiseurs de pluie, il est facile de calculer k_α qui vaut (*cf.* § 14.2.5) :

$$\exp\left(-\frac{9}{2.10^4}[(655-650)^2-(655-600)^2]\right) = \exp(1.35) = 3.857$$

Pour rejeter H_0, il aurait fallu que H_1 soit près de 4 fois plus vraisemblable que H_0.

Dans un contexte bayésien on peut parler de probabilités *a posteriori* (c'est à dire une fois connues les observations) si on s'est donné des probabilités *a priori* sur les états de la nature.

Il faut donc ici se donner π_0 et $\pi_1 = 1 - \pi_0$, probabilités *a priori* de H_0 et H_1 qui quantifient notre information préalable.

Si x désigne le vecteur des observations :

$$P(H_0/\mathbf{x}) = \frac{\pi_0 L_0(\mathbf{x})}{\pi_0 L_0(\mathbf{x}) + \pi_1 L_1(\mathbf{x})}$$

On peut remplacer les vraisemblances par les densités de la variable de décision T, si T est une statistique exhaustive.

La règle bayésienne consiste à choisir l'hypothèse la plus probable *a posteriori*, donc celle qui a une probabilité supérieure à 0.5. On vérifie alors que le test de Neyman-Pearson est en fait un test bayésien avec une probabilité *a priori* implicite que l'on peut calculer aisément en combinant $P(H_1/\mathbf{x}) = \dfrac{(1-\pi_0)L_1(\mathbf{x})}{\pi_0 L_0(\mathbf{x}) + (1-\pi_0)L_1(\mathbf{x})} > 0.5$ et $\dfrac{L_1(\mathbf{x})}{L_0(\mathbf{x})} > k$

Ce qui donne :
$$\pi_0 = \frac{k}{1+k}$$

Pour l'exemple des faiseurs de pluie, on trouve que $\pi_0 = 0.79$ ce qui montre bien encore une fois que H_0 est favorisée.

Lorsque les probabilités *a priori* sont connues, on définit le facteur de Bayes qui est le rapport des « odds ratios » des deux hypothèses :

$$B = \frac{P(H_0/\mathbf{x})/P(H_1/\mathbf{x})}{\pi_0/\pi_1}$$

On peut interpréter B comme mesurant la variation du rapport des chances en faveur de H_0 contre H_1, dûe à la prise en compte des données.

Pour des hypothèses simples, on trouve facilement que B est égal au rapport des vraisemblances $B = \dfrac{L_0(\mathbf{x})}{L_1(\mathbf{x})}$, ce qui réconcilie le point de vue bayésien et le point de vue classique, car B ne dépend pas des probabilités *a priori*.

14.3 TESTS PORTANT SUR UN PARAMÈTRE

14.3.1 Moyenne d'une loi LG(m, σ)

14.3.1.1 σ connu

Le test repose sur la variable de décision $\bar{X}$.

Ainsi pour $H_0 : m = m_0$ contre $H_1 : m = m_1$ avec $m_1 > m_0$, la région critique est définie par $\bar{X} > k$. k se détermine en considérant que $\bar{X}$ suit une $LG\left(m\,;\frac{\sigma}{\sqrt{n}}\right)$:

$$P(\bar{X} > k \mid m_0) = P\left(U > \frac{k - m_0}{\sigma/\sqrt{n}}\right) = \alpha$$

Pour un exemple on se reportera à l'introduction de ce chapitre.

14.3.1.2 σ inconnu

La variable de décision est la variable de Student :

$$T_{n-1} = \frac{\overline{X} - m}{S}\sqrt{n-1}$$

Ainsi pour $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ la région critique est définie par :

$$|T_{n-1}| > k \qquad \text{avec } P(|T_{n-1}| > k) = \alpha$$

$$T_{n-1} = \frac{\overline{X} - m_0}{S}\sqrt{n-1}$$

Exemple : $H_0 : m = 30$ contre $H_1 : m > 30$

Un échantillon de 15 observations a donné $\bar{x} = 37.2$ et $s = 6.2$.

On en déduit $t = \dfrac{37.2 - 30}{6.2}\sqrt{14} = 4.35$.

La valeur critique à $\alpha = 0.05$ (test unilatéral) pour un T_{14} est 1.761 : on rejette H_0.

Remarque sur les tests de moyenne : Si la variable parente ne suit pas une loi de Gauss, les tests précédents s'appliquent encore dès que n est assez grand ($n > 30$ environ) en raison du théorème central-limite.

14.3.2 Variance d'une loi de LG(m, σ)

14.3.2.1 m connu

La variable de décision est $D = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$.

Ainsi pour $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma = \sigma_1$ avec $\sigma_1 > \sigma_0$ la région critique est définie par $\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 > k$ et K est déterminé en considérant que $\frac{nD}{\sigma^2}$ suit un χ^2_n :

$$P(D > k) = P\left(\chi^2_n > \frac{nk}{\sigma_0^2}\right) = \alpha$$

14.3.2.2 m inconnu

La variable de décision est $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ qui est telle que $\frac{nS^2}{\sigma^2}$ suit un χ^2_{n-1}.

Ainsi pour $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma = \sigma_1$ avec $\sigma_1 > \sigma_0$ la région critique est définie par $S^2 > k$ et k est déterminé par :

$$P(S^2 > k) = P\left(\chi^2_{n-1} > \frac{nk}{\sigma_0^2}\right) = \alpha$$

■ **Exemple :** $H_0 : \sigma = 3$ contre $H_1 : \sigma > 3$

Avec 20 observations on a trouvé $s = 3.5$, soit $s^2 = 12.25$.

La valeur critique d'un χ^2_{19} pour $\alpha = 0.05$ est 30.144 d'où :

$$k = \frac{30.144 \times 9}{20} = 13.56$$

La valeur constatée s^2 étant inférieure, on ne peut donc rejeter H_0 au seuil choisi de 0.05. ■

Remarque sur les tests de variance : Les tests précédents utilisant la loi du χ^2 ne sont valables ***que*** dans le cas où ***X suit une loi de Gauss.***

14.3.3 Test de la valeur théorique p d'un pourcentage pour un grand échantillon

On utilise la fréquence empirique F qui suit approximativement une loi :

$$LG\left(p ; \sqrt{\frac{p(1-p)}{n}}\right)$$

$H_0 : p = p_0$ contre $H_1 : p \neq p_0$. La région critique est :

$$|F - p_0| > u_{\alpha/2}\sqrt{p_0\frac{(1-p_0)}{n}}$$

■ **Exemple :** Sur un échantillon de 200 individus d'une commune, 45 % sont favorables à l'implantation d'un centre commercial. Ceci contredit-il l'hypothèse qu'un habitant sur deux y est favorable ?

$$\begin{cases} H_0 : p = 0.5 \\ H_1 : p \neq 0.5 \end{cases}$$

avec $\alpha = 0.05$ $u = 1.96$ d'où la région critique :

$$|F - 0.5| > 1.96\sqrt{\frac{(0.5)^2}{200}}, \text{ soit } W = \{|F - 0.5| > 0.07\}$$

Comme $|f - 0.50| = 0.05$, on ne peut rejeter H_0 au seuil $\alpha = 0.05$.

Si n est trop petit pour que l'on puisse appliquer la loi de Laplace-Gauss, on utilisera l'abaque elliptique (table A.3 *bis*). ■

14.4 TESTS DE COMPARAISON D'ÉCHANTILLONS

14.4.1 Tests de Fisher-Snedecor et de Student pour échantillons indépendants

Étant donné deux échantillons de taille n_1 et n_2, peut-on admettre qu'ils ont été prélevés dans une même population relativement à la variable étudiée, ces deux échantillons ayant été prélevés indépendamment l'un de l'autre ?

Exemples :

- Les résultats scolaires des filles et des garçons sont-ils comparables ?
- Les demandes de deux produits A et B échantillonnées sur un an sont-elles comparables ?

Mathématiquement le problème se formalise de la manière suivante : on observe sur le premier échantillon les réalisations d'une variable aléatoire X_1 de fonction de répartition $F_1(x)$ et sur le deuxième échantillon les réalisations d'une variable aléatoire X_2 de fonction de répartition $F_2(x)$ on veut tester :

$$\begin{cases} H_0 : F_1(x) = F_2(x) \\ H_1 : F_1(x) \neq F_2(x) \end{cases}$$

Le choix de H_0 est dicté par des considérations pratiques car $F_1(x) \neq F_2(x)$ est trop vague pour obtenir une région critique.

Dans la pratique on se contentera de vérifier l'égalité des espérances et des variances de X_1 et X_2, en disposant de $\bar{x}_1$ et $\bar{x}_2$ et s_1^2 et s_2^2 moyennes et variances empiriques des deux échantillons si les lois de X_1 et X_2 sont gaussiennes.

14.4.1.1 *Cas de deux échantillons gaussiens $X_1 \in LG(m_1, \sigma_1)$ et $X_2 \in LG(m_2, \sigma_2)$*

Les hypothèses deviennent alors :

$$H_0 : m_1 = m_2 \text{ et } \sigma_1 = \sigma_2 \quad \text{contre} \quad H_1 : m_1 \neq m_2 \text{ et } \sigma_1 \neq \sigma_2$$

Le test va consister à tester d'abord les variances et si elles ne sont pas significativement différentes à tester ensuite les espérances en admettant $\sigma_1 = \sigma_2$.

- Test des variances de Fisher-Snedecor

En appliquant les résultats de la théorie de l'échantillonnage :

$$\frac{n_1 S_1^2}{\sigma_1^2} \in \chi^2_{n_1-1} \qquad \frac{n_2 S_2^2}{\sigma_2^2} \in \chi^2_{n_2-1}$$

Dans l'hypothèse $H_0 : \sigma_1 = \sigma_2$ et l'on a :

$$\boxed{F_{n_1-1\,;\,n_2-1} = \frac{\dfrac{n_1 S_1^2}{n_1-1}}{\dfrac{n_2 S_2^2}{n_2-1}}}$$

On peut interpréter F comme le rapport des deux estimateurs de σ_1^2 et σ_2^2 respectivement. Si $\sigma_1 = \sigma_2$, ce rapport ne doit pas différer significativement de 1. F sera la variable de décision. En pratique on met toujours au numérateur la plus grande des deux quantités :

$$\frac{n_1 S_1^2}{n_1-1} \quad \text{et} \quad \frac{n_2 S_2^2}{n_2-1}$$

et la région critique est de la forme $F > k$ avec $k > 1$.

Si les deux échantillons ont même taille $n_1 = n_2 = n$, le calcul se simplifie et :

$$F_{n-1,n-1} = \frac{S_1^2}{S_2^2}$$

Si le test de Fisher-Snedecor aboutit à la conclusion $\sigma_1 = \sigma_2$, on passe au test des espérances.

Exemple :

$$n_1 = 25, \quad n_2 = 13, \quad s_1^2 = 0.05, \quad s_2^2 = 0.07, \quad \alpha = 0.05$$

Il faut permuter les indices 1 et 2 car $\frac{13 \times 0.07}{12} > \frac{25 \times 0.09}{24}$.

La région critique est $F > 2.18$.

On accepte l'hypothèse $\sigma_1 = \sigma_2$.

- Test des espérances de Student

Supposons désormais $\sigma_1 = \sigma_2 = \sigma$. On a :

$$\begin{cases} \dfrac{n_1 S_1^2}{\sigma^2} \in \chi^2_{n_1-1} \\ \overline{X}_1 \in LG\left(m_1, \dfrac{\sigma}{\sqrt{n_1}}\right) \end{cases} \quad \text{et} \quad \begin{cases} \dfrac{n_2 S_2^2}{\sigma^2} \in \chi^2_{n_2-1} \\ \overline{X}_2 \in LG\left(m_2, \dfrac{\sigma}{\sqrt{n_2}}\right) \end{cases}$$

d'où :

$$\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \in \chi^2_{n_1+n_2-2}$$

et :

$$\overline{X}_1 - \overline{X}_2 \in LG\left(m_1 - m_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

σ étant inconnu on utilise la loi de Student.

Par définition de la variable de Student :

$$T_{n_1+n_2-2} = \frac{\dfrac{\overline{X}_1 - \overline{X}_2 - (m_1 - m_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\sqrt{\dfrac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2(n_1 + n_2 - 2)}}}$$

Ce qui se simplifie en éliminant σ :

$$\boxed{T_{n_1+n_2-2} = \frac{(\overline{X}_1 - \overline{X}_2) - (m_1 - m_2)}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}\sqrt{n_1 + n_2 - 2}}$$

Dans l'hypothèse H_0, $m_1 = m_2$ et la région critique est de la forme : $|T| > k$.

On aura vu au passage que seule l'hypothèse H_0 d'égalité des moyennes et des variances permet d'obtenir des régions critiques, car on élimine précisément les valeurs communes de ces moyennes et variances.

De plus l'ordre : test de variances, puis test de moyennes, semble indispensable, car le test de Student suppose explicitement $\sigma_1 = \sigma_2$.

14.4.1.2 Comparaison de moyennes en cas de variances inégales

Lorsque les effectifs des deux échantillons sont élevés (supérieurs chacun à 20), la formule précédente reste encore approximativement valable.

Pour de petits échantillons, l'approximation d'Aspin-Welch est souvent utilisée dans les logiciels : elle consiste à remplacer le degré de liberté $n_1 + n_2 - 2$, par une valeur inférieure m obtenue par les formules suivantes :

$$m = \frac{1}{\dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}} \quad \text{avec } c = \frac{\dfrac{S_1^2}{n_1 - 1}}{\dfrac{S_1^2}{n_1 - 1} + \dfrac{S_2^2}{n_2 - 1}}$$

14.4.1.3 Cas d'échantillons non gaussiens

Le test de variance F n'est plus valable car $\dfrac{nS^2}{\sigma^2}$ ne suit pas une loi de χ^2, mais on a le résultat suivant qui permet de tester $m_1 = m_2$.

Pour n_1, n_2 assez grand (quelques dizaines d'observations) on peut quand même tester les moyennes m_1 et m_2 en appliquant la formule de Student ***que σ_1 soit différent ou non de σ_2.***

On dit que le test de Student est « robuste » car il résiste bien à un changement de la loi de X_1 et X_2.

14.4.2 Tests non paramétriques de comparaison de deux échantillons indépendants

14.4.2.1 Test de Smirnov

Ce test est analogue au test de Kolmogorov et repose sur le résultat suivant.

Soit $F^*_{n_1}(x)$ et $F^*_{n_2}(x)$ les fonctions de répartition empiriques de deux échantillons de taille n_1 et n_2 issues d'une même loi, de fonction de répartition $F(x)$; alors :

$$P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_x |F^*_{n_1}(x) - F^*_{n_2}(x)| < y\right) \to K(y)$$

D'où le test : soit à tester l'hypothèse $H_0 F(x) = G(x)$, contre $H_1 F(x) \neq G(x)$, en disposant de deux échantillons de taille n_1 et n_2 de fonctions de répartition empiriques $F^*_{n_1}(x)$ et $G^*_{n_1}(x)$, on forme la différence des deux et on en prend le sup et on rejette H_0 si sup $|F^*_{n_1}(x) - G^*_{n_2}(x)|$ est trop grand.

14.4.2.2 Test de Wilcoxon-Mann-Whitney

Soit $(x_1, x_2, \ldots, x_n)$ et $(y_1, y_2, \ldots, y_m)$ les deux échantillons. Ce test repose sur l'idée que si l'on mélange les deux séries de valeurs et qu'on ordonne le tout par valeurs croissantes on doit obtenir un mélange homogène.

Pour cela les deux suites étant réordonnées, on compte le nombre total de couples (x_i, y_i) où x_i a un rang plus grand que y_i (ou bien tels que $x_i > y_i$ si X et Y sont quantitatives).

Soit U ce nombre (statistique de Mann-Whitney). Il est facile de voir que U varie de 0 à nm ; si $U = 0$ on a la situation suivante (mélange en deux phases) :

$$x_1, x_2, \ldots, x_n, \ y_1, y_2, \ldots, y_m$$

de même si $U = nm$:

$$y_1, y_2, \ldots, y_m, \ x_1, x_2, \ldots, x_n$$

D'autre part si les deux distributions sont issues de la même population :

$$E(U) = \frac{nm}{2} \quad \text{et} \quad V(U) = \frac{nm(n + m + 1)}{12}$$

et asymptotiquement U est gaussien, l'approximation étant excellente dès que n et m sont supérieurs ou égaux à 8. Dans tous les cas on peut calculer la loi exacte de U.

Le test consistera donc à rejeter $H_0 : F(x) = G(x)$ si $\left| U - \frac{nm}{2} \right| > k$.

Un autre mode de calcul plus rapide dans certain cas consiste à calculer la somme des rangs des individus de l'un des deux groupes (le premier par exemple).

Soit W_X cette somme appelée statistique de Wilcoxon. Il est facile de montrer que $W_X = nm + \frac{n(n + 1)}{2} - U$ sous l'hypothèse nulle :

$$E(W_X) = \frac{n(n + m + 1)}{2}$$

$$V(W_X) = \frac{nm(n + m + 1)}{12}$$

La région critique est alors définie si n et $m > 8$ par :

$$\left| W_X - \frac{n(n + m + 1)}{12} \right| > u_{\alpha/2} \sqrt{\frac{nm(n + m + 1)}{12}}$$

Exemple : On veut comparer les performances de deux groupes d'élèves à des tests d'habileté manuelle.

On choisit aléatoirement 8 individus du premier groupe et 10 du deuxième. Les performances en minutes sont les suivantes :

Groupe 1 :	22	31	14	19	24	28	27	28		
Groupe 2 :	25	13	20	11	23	16	21	18	17	26

On réordonne les 18 observations par ordre croissant. Les résultats du premier groupe sont soulignés :

Observations :	11	13	<u>14</u>	16	17	18	<u>19</u>	20	21	<u>22</u>	23	<u>24</u>	25	26	<u>27</u>	<u>28</u>	<u>28</u>	<u>31</u>
Rangs :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

La somme des rangs des individus du premier groupe est :

$$W_X = 3 + 7 + 10 + 12 + 15 + 16 + 17 + 18 = 98$$

Si H_0 était vraie :

$$E(W_X) = \frac{8(8+10+1)}{2} = 76 \qquad V(W_X) = \frac{8 \times 10(8+10+1)}{12} = 126.7 = (11.25)^2$$

Comme $\frac{98-76}{11.25} = 1.96$, on peut rejeter H_0 avec $\alpha = 0.10$ et conclure à une plus grande rapidité des élèves du groupe 2.

Remarque :

$$\bar{x}_1 = 24.13 \quad \text{et} \quad \bar{x}_2 = 19$$
$$s_1^2 = 27.36 \quad \text{et} \quad s_2^2 = 22$$

Le test de Fisher-Snedecor de comparaison des variances donne :

$$f = \frac{\frac{27.36 \times 8}{7}}{\frac{22 \times 10}{9}} = 1.28$$

ce qui montre que σ_1 n'est pas significativement différent de σ_2 ($F_{0.05}(7\,;9) = 3.29$).

Le test de Student de différence des moyennes donne :

$$t_{16} = \frac{24.13 - 19}{\sqrt{\left(\frac{1}{10} + \frac{1}{8}\right)(10 \times 22 + 8 \times 27.36)}}\sqrt{16} = 2.06$$

ce qui est supérieur au seuil à 0.10 qui vaut 1.745. Le test de Student conduit donc à la même conclusion que le test de Wilcoxon. Cependant ici, rien ne justifiant l'hypothèse de distributions gaussiennes et vu la petite taille des échantillons, seul le test de Wilcoxon est d'usage légitime.

14.4.3 Test non paramétrique de comparaison de plusieurs échantillons décrits par une variable qualitative : le test du χ^2

Les données se présentent sous la forme du tableau 14.2 :

TABLEAU 14.2

	Modalité 1	Modalité 2		Modalité r	Total
Échantillon 1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
Échantillon 2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
⋮					
Échantillon k	n_{k1}	n_{k2}		n_{kr}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

où n_{ij} est le nombre des individus de l'échantillon possédant la modalité j de la variable :

$$n_{i.} = \sum_{j=1}^{r} n_{ij} = \text{effectif de l'échantillon } i\ ;$$

$$n_{.j} = \sum_{i=1}^{k} n_{ij} = \text{nombre total des individus possédant } j\ ;$$

$$n = \sum_i \sum_j n_{ij} = \sum_i n_{i.} = \sum_j n_{.j}.$$

Il s'agit de tester H_0 : « les échantillons proviennent de la même population » contre H_1 : « les échantillons sont significativement différents ».

Dans l'hypothèse H_0 on peut parler de probabilités $p_1, p_2, \ldots, p_r$ de posséder les modalités 1, 2, . . ., r. Il s'agit alors de comparer les effectifs constatés n_{ij} aux effectifs espérés $n_{i.}\,p_j$, qui ne doivent pas en différer beaucoup ; on forme alors :

$$d_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - n_{i.}\,p_j)^2}{n_{i.}\,p_j}$$

Dans l'hypothèse H_0, d_0^2 est une réalisation d'une variable D_0^2 suivant un χ^2 dont nous allons chercher le nombre de degrés de liberté.

d_0^2 porte sur kr termes, mais ces kr termes sont liés par k relations qui indiquent que les sommes de lignes sont constantes $\sum_j n_{ij} = \sum_j n_{i.}\,p_j = n_{i.}$

Donc D_0^2 est un χ^2_{kr-k}.

Cependant en pratique les $p_1, p_2, \ldots, p_r$ sont rarement connus, et on les estime par $\hat{p}_j = \frac{n_{.j}}{n}$, ce qui fait $r - 1$ estimations indépendantes (en effet pour estimer les r probabilités on n'a besoin que de $r - 1$ relations car $\sum_{j=1}^{r} p_j = 1$).

D'où une nouvelle mesure :

$$d^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i.}\,n_{.j}}{n}\right)^2}{\frac{n_{i.}\,n_{.j}}{n}} = \left(\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{n_{ij}^2}{n_{i.}\,n_{.j}} - 1\right)n$$

alors D^2 est un : $\chi^2_{kr-k-(r-1)} = \chi^2_{(k-1)(r-1)}$ si H_0 est vraie.

On peut remarquer que si l'on utilise la fréquence $f_j = \frac{n_{ij}}{n}$:

$$d^2 = n\sum_i \sum_j \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

Le test consistera donc à rejeter H_0 si d^2 constaté est trop grand, comme pour un test habituel du χ^2.

14.4.4 Test de comparaison de deux pourcentages (grands échantillons)

Dans deux échantillons de grandes tailles n_1 et n_2, on relève les pourcentages f_1 et f_2 d'individus présentant un certain caractère. Soit p_1 et p_2 les probabilités correspondantes : il s'agit de savoir si f_1 et f_2 sont significativement différents ou non, donc de tester :

$$\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 \neq p_2 \end{cases}$$

Si H_0 est vraie, f_1 et f_2 sont des réalisations indépendantes de deux variables F_1 et F_2 suivant les lois :

$$LG\left(p\,;\sqrt{\frac{p(1-p)}{n_1}}\right) \quad \text{et} \quad LG\left(p\,;\sqrt{\frac{p(1-p)}{n_2}}\right)$$

donc :

$$F_1 - F_2 LG\left(0\,;\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

On rejettera H_0, si, avec $\alpha = 0.05$ par exemple :

$$|f_1 - f_2| > 1.96\sqrt{p(1-p)}\sqrt{\frac{1}{n_2} + \frac{1}{n_2}}$$

Lorsque p n'est pas connu on le remplace par son estimation $\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$.

Exemple :

- sur 96 pièces venant d'un fournisseur A, 12 sont défectueuses ;
- sur 55 pièces venant d'un fournisseur B, 15 sont défectueuses.

Les pourcentages de pièces défectueuses sont-ils significativement différents ?

$$f_1 = 0.13 \qquad f_2 = 0.27 \qquad \hat{p} = \frac{12 + 15}{96 + 55} = 0.18$$

$$\frac{f_1 - f_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -2.28$$

On peut donc rejeter l'hypothèse $H_0 : p_1 = p_2$ et conclure que $p_2 > p_1$.

Le même test aurait pu être réalisé de manière équivalente en utilisant le test du χ^2 de comparaison d'échantillons. Le calcul est d'ailleurs particulièrement simple dans le cas du tableau à quatre cases (voir chapitre 6 § 6.5.2.2) :

	Défectueux	Non défectueux	
Fournisseur A	12	84	96
Fournisseur B	15	40	55
	27	124	151

$$d^2 = \frac{151(12 \times 40 - 15 \times 84)^2}{27 \times 124 \times 96 \times 55} = 5.20$$

Avec un degré de liberté la valeur critique du χ^2 pour $\alpha = 0.05$ est 3.84 ; on rejette donc H_0.

On aura remarqué que $5.20 = (2.28)^2$ car (la démonstration est laissée au soin du lecteur) on a exactement :

$$d^2 = \left(\frac{f_1 - f_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right)^2$$

14.4.5 Comparaison des moyennes de deux échantillons gaussiens indépendants à p dimensions de même matrice de variance

Considérons deux échantillons de n_1 et n_2 observations issus respectivement de deux lois $N_p(\boldsymbol{\mu}_1 ; \boldsymbol{\Sigma})$ et $N_p(\boldsymbol{\mu}_2 ; \boldsymbol{\Sigma})$. On cherche alors à tester :

$$\begin{cases} H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \\ H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \end{cases}$$

On notera $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{V}_1$, $\mathbf{V}_2$, les centres de gravité et matrices de variance des deux échantillons, et on posera $\mathbf{W} = \frac{n_1\mathbf{V}_1 + n_2\mathbf{V}_2}{n_1 + n_2}$ la matrice de variance intragroupe $\left(\frac{n_1 + n_2}{n_1 + n_2 - 2}\mathbf{W}\text{ est un estimateur sans biais de }\boldsymbol{\Sigma}\right)$.

14.4.5.1 Test de Hotelling

$n_1\mathbf{V}_1 + n_2\mathbf{V}_2$ est une matrice de Wishart $W_p(n_1 + n_2 - 2\ ;\ \boldsymbol{\Sigma})$ et $\mathbf{g}_1 - \mathbf{g}_2$ une loi $N_p\left(\mathbf{0}\ ;\boldsymbol{\Sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$ si H_0 est vraie.

On en déduit (chapitre 4, paragr. 4.5) :

$$\frac{n_1 n_2(n_1 + n_2 - 2)}{(n_1 + n_2)^2}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = T_p^2(n_1 + n_2 - 2)$$

si H_0 est vraie d'où le test.

En pratique on utilisera la relation entre T_p^2 et F qui donne :

$$\frac{(n_1 + n_2 - p - 1)n_1 n_2}{p(n_1 + n_2)^2}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = F(p\ ; n_1 + n_2 - p - 1)$$

On vérifiera que pour $p = 1$ on retrouve le carré de la variable de Student du test décrit au paragraphe 14.5.1 de ce chapitre.

14.4.5.2 Distance de Mahalanobis

Le test précédent est plus couramment présenté sous la forme suivante.

Soit $\Delta_p^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ le carré de la distance de Mahalanobis entre $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$. Le test revient donc à poser :

$$\begin{cases} H_0 : \Delta_p^2 = 0 \\ H_1 : \Delta_p^2 > 0 \end{cases}$$

La distance de Mahalanobis estimée D_p est telle que :

$$D_p^2 = \frac{n_1 + n_2 - 2}{n_1 + n_2}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

obtenue en remplaçant $\boldsymbol{\Sigma}$ par son estimation sans biais. Remarquons que ceci ne revient pas à estimer sans biais $\boldsymbol{\Sigma}^{-1}$ et que :

$$E(D_p^2) = \frac{n_1 + n_2 - 2}{n_1 + n_2 - p - 1}\left(\Delta_p^2 + p\frac{n_1 + n_2}{n_1 n_2}\right) > \Delta_p^2$$

Lorsque $\Delta_p^2 = 0$, $\frac{n_1 n_2}{n_1 + n_2} D_p^2$ suit un $T_p^2(n_1 + n_2 - 2)$ d'où le résultat :

$$\boxed{\frac{n_1 n_2}{(n_1 + n_2)}\frac{(n_1 + n_2 - p - 1)}{p(n_1 + n_2 - 2)} D_p^2 = F(p\ ; n_1 + n_2 - p - 1)}$$

Cette expression est couramment appliquée en analyse discriminante (voir chapitre 18).

14.4.6 Comparaison de moyennes d'échantillons appariés

Un même échantillon d'individus est soumis à deux mesures successives d'une « même » variable.

14.4.6.1 Le cas gaussien

Exemples :

- 15 copies sont soumises à une double correction ;
- passage du même test d'aptitude à deux instants différents (problème de l'apprentissage).

On veut tester l'hypothèse que les deux séries de valeurs sont semblables. Soit X_1 la variable correspondant à la première série et X_2 l'autre. En fait on se contente de tester l'hypothèse $E(X_1) = E(X_2)$ en posant le modèle suivant :

$$X_1 - X_2 \, LG(m_1 - m_2, \sigma)$$

(ce qui sous-entend que X_1 et X_2 sont séparément gaussiens).

Le test de $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$ consiste à former les différences $x_{i1} - x_{i2} = d_i$ et à faire un test de Student sur la moyenne des d_i car σ est en général inconnu :

$$t_{n-1} = \frac{\bar{d}}{s_d}\sqrt{n-1} = \frac{\bar{X}_1 - \bar{X}_2}{s_d}\sqrt{n-1}$$

On rejettera H_0 si $|t| > k$.

N.B. **:** La différence avec le test de Student d'égalité de deux moyennes étudié au paragraphe 14.1.1 provient du fait que les variables X_1 et X_2 ne peuvent ici être supposées indépendantes : la variance de leur différence ne peut être estimée par la somme des variances.

Exemple : Considérons deux séries de mesures effectuées sur les mêmes individus à deux instants différents

Individu	X_1	X_2	$D = X_1 - X_2$
1	86	66	20
2	92	76	16
3	75	63	12
4	84	62	22
5	66	74	−8
6	75	70	5
7	97	86	11
8	67	69	−2
9	99	81	18
10	68	92	−24

Les moyennes ont-elles varié ?

On trouve $\bar{d} = 7 \quad s^* = 14.56 \qquad t = \frac{7}{14.56/\sqrt{10}} = 1.52$

On ne peut donc rejeter l'hypothèse que les deux moyennes sont égales car la valeur critique d'un test bilatéral à 5 % vaut 2.269 pour un T_9.

Le test précédent suppose la normalité des deux variables. Si ce n'est pas le cas, ou si cette hypothèse ne peut être prouvée, il peut être plus prudent, mais avec une puissance moindre, d'effectuer un test non paramétrique.

14.4.6.2 Test des signes

On compte le nombre K de différences positives. Sous l'hypothèse nulle d'absence de différence entre moyennes, il y a une chance sur deux qu'une différence soit positive ou négative ; donc K suit une loi binomiale $\mathcal{B}(10 ; 0.5)$. Dans l'exemple il y a 7 différences positives. Or $P(K < 8) = 0.9453$. Avec un test bilatéral à 5 %, la conclusion reste alors la même.

14.4.6.3 Le test de Wilcoxon pour données appariées

Il est bien plus puissant que le test des signes et doit lui être préféré. Il teste en réalité une hypothèse alternative de distribution décalée.

On procéde comme suit : on classe par ordre de valeurs absolues croissantes les différences :

Rang	*D*
1	−24
2	22
3	20
4	18
5	16
6	12
7	11
8	−8
9	5
10	−2

On calcule ensuite la somme des rangs des différences positives soit ici :

$$W_+ = 2 + 3 + 4 + 5 + 6 + 7 + 9 = 36$$

Sous l'hypothèse nulle, on trouve aisément l'espérance et la variance de W_+

En effet $W_+ = \sum_{i=1}^{n} R_i Z_i$ où $\begin{cases} Z_i = 1 \text{ si } X_{1i} - X_{2i} > 0 \\ Z_i = 0 \text{ sinon} \end{cases}$ en ne tenant pas compte des ex-aequo. Les R_i sont les rangs de toutes les différences et sont donc une permutation des entiers de 1 à n.

Les Z_i sont des variables de Bernoulli indépendantes :

$$E(Z_i) = \frac{1}{2} \quad V(Z_i) = \frac{1}{4}$$

$$W_+ = \sum_{i=1}^{n} R_i Z_i$$

conditionnellement aux rangs :

$$E(W_+/R) = \sum_{i=1}^{n} r_i E(Z_i) = \frac{1}{2}\sum_{i=1}^{n} r_i = \frac{1}{2}\sum_{i=1}^{n} i = \frac{1}{2}\frac{n(n+1)}{2} = \frac{n(n+1)}{4}$$

$$V(W_+/R) = \sum_{i=1}^{n} r_i^2\, V(Z_i) = \frac{1}{4}\sum_{i=1}^{n} r_i^2 = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{1}{4}\frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}$$

Les rangs ayant disparus, l'espérance et la variance conditionnelle sont l'espérance et la variance totale.

On peut montrer que W_+ peut être approximé par une loi normale à partir de $n = 10$.

On comparera donc la valeur trouvée 36 à une loi normale d'espérance 27.5 et de variance 96.25, ce qui donne une valeur centrée réduite de 0.87. On ne rejette pas l'hypothése nulle.

14.4.7 Comparaison de variances d'échantillons appariés

Les variables n'étant pas indépendantes, le test de Fisher-Snedecor ne peut être appliqué.

On utilisera la propriété suivante :

$$\text{cov}(X_1 + X_2 ; X_1 - X_2) = V(X_1) - V(X_2)$$

Tester l'égalité des variances revient donc à tester si le coefficient de corrélation linéaire entre la somme et la différence des deux variables est nul.

Dans l'exemple précédent on trouve $r = 0.224$ ce qui ne permet pas de rejeter l'hypothèse d'égalité des variances car le seuil à 5 % bilatéral pour 10 observations est 0.63 (*cf.* table A.9).

Les écart-types corrigés étaient respectivement 12.45 et 9.95

14.4.8 Le test de Mc Nemar de comparaison de deux pourcentages pour un même échantillon

On a vu en 14.4.4 comment comparer des pourcentages sur deux échantillons indépendants. Qu'en est-il lorsqu'il s'agit des mêmes individus ? Par exemple, on interroge à deux reprises, après une action, 600 clients d'une société pour connaître leur taux de satisfaction.

On commettrait une grave erreur en appliquant les formules des échantillons indépendants : il faut ici connaître pour chaque individu son état aux deux enquêtes, que l'on peut résumer par le tableau de contingence 2 × 2 croisant les effectifs des deux variables.

Prenons l'exemple suivant :

T_1 \ T_2	**Satisfaits**	**Non satisfaits**
Satisfaits	200	50
Non satisfaits	80	270

La proportion de satisfaits est passée de 41.7 % à 46.7 %. S'il s'agissait de deux échantillons indépendants de 600 individus, cette différence ne serait pas jugée significative.

Mais pour tester la significativité de cette différence, il faut en réalité comparer les effectifs des individus ayant changé d'avis.

En effet, avec des notations classiques, l'hypothèse H_0 est $p_{1.} = p_{.1}$. Comme $p_{1.} = p_{11} + p_{12}$ et $p_{.1} = p_{11} + p_{21}$ H_0 revient à tester $p_{12} = p_{21}$

T_1 \ T_2	Satisfaits	Non satisfaits	
Satisfaits	p_{11}	p_{12}	$p_{1.}$
Non satisfaits	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	

On effectue alors un classique test du khi-deux : sous l'hypothèse nulle $p_{12} = p_{21}$ est estimé par $(n_{12} + n_{21})/2$.

La statistique de test est donc :

$$\frac{\left(n_{12} - \dfrac{n_{12} + n_{21}}{2}\right)^2 + \left(n_{21} - \dfrac{n_{12} + n_{21}}{2}\right)^2}{\dfrac{n_{12} + n_{21}}{2}}$$

Un calcul facile montre qu'elle est égale à :

$\dfrac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$. On comparera cette quantité à un χ^2_1, ou sa racine carrée à une variable normale centrée réduite.

Ici on trouve $\dfrac{(n_{12} - n_{21})}{\sqrt{n_{12} + n_{21}}} = \dfrac{80 - 50}{\sqrt{80 + 50}} = 2.63$. On conclue à une augmentation significative de la satisfaction.

14.5 L'ANALYSE DE VARIANCE

L'analyse de variance recouvre un ensemble de technique de tests et d'estimation destinés à apprécier l'effet de variables qualitatives sur une variable numérique et revient dans le cas simple à comparer plusieurs moyennes d'échantillons gaussiens.

On utilisera ici un vocabulaire particulier : les variables qualitatives susceptibles d'influer sur la distribution de la variable numérique observée sont appelées « ***facteurs de variabilité*** » et leurs modalités « ***niveaux*** ». Lorsqu'il y a plusieurs facteurs, une combinaison de niveaux est un « ***traitement*** » (voir chapitre 21).

Le domaine étant très vaste on se contentera ici d'une brève présentation du modèle à effets fixes à un et deux facteurs (des compléments seront donnés au chapitre 17 sur le modèle linéaire général).

14.5.1 Analyse de variance à un facteur

14.5.1.1 Les données et le modèle

On dispose de k échantillons de tailles respectives $n_1, n_2, \dots, n_k$ correspondant chacun à un niveau différent d'un facteur A. On pose $n = \sum_{i=1} n_i$ et on dresse le tableau 14.3.

On suppose que le facteur A influe uniquement sur les moyennes des distributions et non sur leur variance. Il s'agit donc d'un test de confusion des k moyennes $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$.

TABLEAU 14.3

Facteur	A_1	A_2		A_i		A_k
	x_1^1	x_2^1	....	x_i^1	....	x_k^1
	x_1^2	x_2^2	....	x_i^2	....	
	$x_1^{n_1}$	$x_2^{n_2}$	....	....	....	$x_k^{n_k}$
Moyennes	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_i$		$\bar{x}_k$

Si on considère chaque échantillon comme issu d'une variable aléatoire X_i suivant une loi LG(m_i ; σ), le problème est donc de tester :

$$\begin{cases} H_0 \ : \ m_1 = m_2 = \dots = m_k = m \\ H_1 \ : \ \exists i, j \ m_i \neq m_j \end{cases}$$

On peut également poser :

$$x_i^j = m_i + \varepsilon_i^j \quad \text{où} \quad \varepsilon_i^j \ \text{LG}(0 \,;\, \sigma)$$

ou encore $x_i^j = \mu + \alpha_i + \varepsilon_i^j$ où μ représente une valeur moyenne et α_i l'effet du niveau i du facteur.

Si H_0 est rejetée le problème se posera donc d'estimer m_i (ou μ et les α_i).

14.5.1.2 Le test

Si $\bar{X}$ est la moyenne totale $\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_i^j$ et en remarquant que :

$$X_i^j - \bar{X} = X_i^j - \bar{X}_i + \bar{X}_i - \bar{X}$$

il vient facilement :

$$\frac{1}{n}\sum_i\sum_j(X_i^j - \bar{X})^2 = \frac{1}{n}\sum_i\sum_j(X_i^j - \bar{X}_i)^2 + \frac{1}{n}\sum_i n_i(\bar{X}_i - \bar{X})^2$$

formule qui n'est autre que celle de la variance totale décomposée en moyenne des variances et variance des moyennes.

Si on pose : $$S^2 = \frac{1}{n}\sum_i\sum_j(X_i^j - \overline{X})^2, \quad S_A^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2$$

et : $$S_R^2 = \frac{1}{n}\sum_i\sum_j(X_i^j - \overline{X}_i)^2$$

on a donc $S^2 = S_A^2 + S_R^2$ formule « d'analyse de variance ».

S_A^2 représente la variance due au facteur, S_R^2 la variance résiduelle.

Si on écrit $S_R^2 = \frac{1}{n}\sum_{i=1}^{k} n_i S_i^2$ avec $S_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_i^j - \overline{X}_i)^2$ en introduisant les dispersions de chaque échantillon, on trouve que $\frac{nS_R^2}{\sigma^2}$ est une variable de χ^2 à $n - k$ degrés de liberté car $\frac{n_i S_i^2}{\sigma^2}$ est une variable $\chi^2_{n_i-1}$ et $\frac{nS_R^2}{\sigma^2} = \sum_{i=1}^{k}\frac{n_i S_i^2}{\sigma^2}$.

Dans l'hypothèse H_0 et dans celle-ci seulement, les X_i sont des variables de même loi, dans ce cas $\frac{nS^2}{\sigma^2}$ suit un χ^2_{n-1} et $\frac{nS_A^2}{\sigma^2}$ un χ^2_{k-1} car S^2 est alors la variance d'un n-échantillon d'une LG(m, σ) et S_A^2 est analogue à la variance du k-échantillon des moyennes $\overline{X}_i$ pondérées par les n_i.

L'équation d'analyse de variance n'est autre que la formule du théorème de Cochran, ce qui entraîne que S_R^2 et S_A^2 sont des variables aléatoires indépendantes, si H_0 est vraie ; on a en effet :

$$\chi^2_{n-1} = \chi^2_{k-1} + \chi^2_{n-k}$$

Donc si H_0 est vraie :

$$\frac{S_A^2/k - 1}{S_R^2/n - k} = F(k - 1\,; n - k)$$

d'où le test : on forme le rapport $\frac{S_A^2/k - 1}{S_R^2/n - k}$. S'il est supérieur à la valeur critique d'une variable de Fisher-Snedecor on conclut à une infuence significative du facteur A.

Le carré moyen résiduel est alors un estimateur sans biais de σ^2.

Exemple : Reprenons l'exemple étudié aux chapitres 5 et 6 : les variations du taux de taxe d'habitation de 100 villes françaises et étudions s'il existe des différences entre zones géographiques. La première étape avait été de comparer les diagrammes en boîte qui montraient des différences essentiellement entre le Nord et l'Ile-de-France et les autres zones.

Le tableau suivant donne les moyennes et variances corrigées par zone :

Zone Géographique	Effectif	Moyenne	Variance
Centre	13	18.1154	3.63619
Est	10	17.662	4.38986
Ile-de-France	26	11.7646	15.0492
Nord	9	25.9511	50.4071
Ouest	14	18.8964	9.59955
Sud-Est	18	19.7694	8.63498
Sud-Ouest	10	20.511	20.6971
Total	100	17.7707	30.5765

Le tableau d'analyse de la variance est alors :

Analyse de variance

Source	Somme des carrés	Ddl	Carré moyen	F	Proba.
Inter-groupes	1706.58	6	284.43	20.03	0.0000
Intra-groupes	1320.49	93	14.1988		
Total (Corr.)	3027.07	99			

On rejette donc l'hypothèse d'égalité des 7 moyennes car la valeur de la statistique de test F est très élevée et la probabilité qu'un $F_{6\,;\,93}$ dépasse 20 est pratiquement nulle. ■

14.5.1.3 L'estimation des effets

Il est clair que les m_i sont estimables par les moyennes $\bar{x}_i$ mais que les α_i ne le sont pas: il y a indétermination puisque $m_i = \mu + \alpha_i$ peut s'obtenir d'une infinité de manières.

On pose généralement la contrainte suivante d'effet moyen nul : $\sum_{i=1}^{k} n_i\alpha_i = 0$ d'où :

$$\hat{\mu} = \bar{x}$$
$$\hat{\alpha}_i = \bar{x}_i - \bar{x}$$

14.5.1.4 Comparaisons multiples de moyennes

Le rejet de H_0 ne signifie pas que tous les m_i sont différents entre eux, et on cherche souvent à tester l'égalité à 0 des différences $m_i - m_j$ (appelées « contrastes »). Diverses méthodes existent.

Un résultat dû à Scheffé montre que pour tout contraste l'événement :

$$m_i - m_j - S\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \leq \bar{x}_i - \bar{x}_j \leq m_i - m_j + S\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

a lieu avec une probabilité $1 - \alpha$ donnée par :

$$P\left(F_{k-1\,;\,n-k} \leq \frac{S^2}{k-1}\right) = 1 - \alpha$$

où $\hat{\sigma}^2$ est le carré moyen résiduel. On rejette H_0 s'il existe au moins un contraste significativement différent de 0.

On peut donc tester simultanément tous les contrastes de la façon suivante : on calcule tout d'abord :

$$S = \sqrt{(k-1)\,F_\alpha(k-1\,;n-k)}$$

et on vérifie ensuite si $|\bar{x}_i - \bar{x}_j| > S\hat{\sigma}\sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_j}}$; si oui : $m_i \neq m_j$.

On prendra garde toutefois que les comparaisons par paires ne sont pas transitives.

On pourrait ainsi accepter $m_1 = m_2$, $m_2 = m_4$, mais pas $m_1 = m_4$!

Il est souvent plus simple de représenter graphiquement les intervalles de confiance déduits de la méthode de Scheffé.

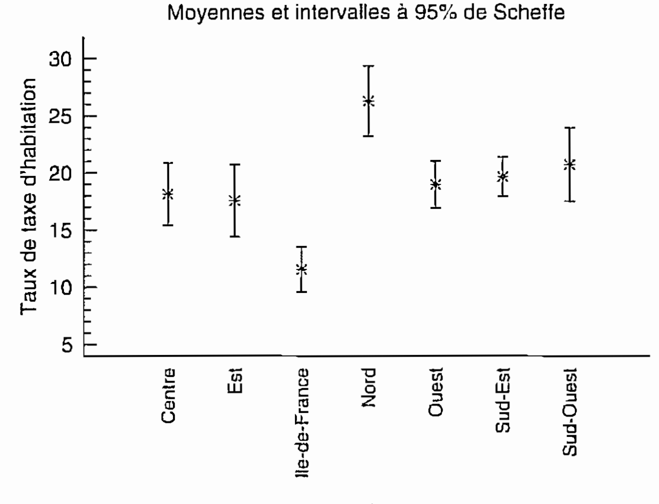


FIGURE 14.5

14.5.1.5 Test de comparaison de k variances

L'hypothèse d'égalité des variances pour chaque niveau joue un rôle important en analyse de variance mais n'est que rarement testée : en effet les tests disponibles sont peu fiables et il vaut mieux se contenter de procédures empiriques.

Citons cependant le test de Bartlett :

Soient $S_1^{*2}, S_2^{*2}, \ldots, S_k^{*2}$ les variances corrigées des k échantillons, si $\sigma_1 = \sigma_2 = \ldots \sigma_k$ alors la quantité :

$$(n-k)\ln\left(\frac{\sum_{i=1}^{k}(n_i-1)\,S_i^{*2}}{n-k}\right) - \sum_{i=1}^{k}(n_i-1)\ln(S_i^{*2})$$

suit approximativement une loi du χ^2 à $k-1$ degrés de liberté.

14.5.2 Analyse de variance à deux facteurs

14.5.2.1 Le modèle

On notera p et q les nombres de niveaux de deux facteurs A et B.

Pour chaque couple i, j de niveaux (traitement) on aura n_{ij} observations de la variable X.

On dit que le modèle est ***complet*** si $n_{ij} > 0$ pour tout traitement, ***à répétition*** si $n_{ij} > 1$, ***équilibré*** si $n_{ij} = r$.

On limitera cette étude au cas équilibré. Les données recueillies sont donc, pour un traitement (i, j), x_{ijk} avec $k = 1, 2, \ldots, r$.

On supposera que x_{ijk} soit une loi LG(m_{ij} ; σ) donc que $x_{ijk} = m_{ij} + \varepsilon_{ijk}$ où ε_{ijk} LG(0 ; σ).

On écrit alors :

$$m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

μ est l'effet moyen ;

α_i l'effet principal du niveau i de A ;

β_j l'effet principal du niveau j de B ;

γ_{ij} l'effet d'interaction.

La présence d'un terme d'interaction équivaut à la non-additivité des effets principaux.

On posera :

$$\bar{x}_{...} = \frac{1}{pqr}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r} x_{ijk}$$

$$\bar{x}_{ij.} = \frac{1}{r}\sum_{k=1}^{r} x_{ijk}$$

$$\bar{x}_{i..} = \frac{1}{qr}\sum_{j=1}^{q}\sum_{k=1}^{r} x_{ijk}$$

$$\bar{x}_{.j.} = \frac{1}{pr}\sum_{i=1}^{p}\sum_{k=1}^{r} x_{ijk}$$

On a alors :

$$\begin{aligned} x_{ijk} - \bar{x}_{...} &= (\bar{x}_{i..} - \bar{x}_{...}) + (\bar{x}_{.j.} - \bar{x}_{...}) \\ &+ (\bar{x}_{ij.} - \bar{x}_{.j.}) - (\bar{x}_{i..} + \bar{x}_{...}) \\ &+ (x_{ijk} - \bar{x}_{ij.}) \end{aligned}$$

les différents termes de cette somme correspondant respectivement aux effets principaux, à l'interaction et à une fluctuation aléatoire.

14.5.2.2 L'équation d'analyse de variance et le test

On vérifie que pour le modèle équilibré on a :

$$\sum_i\sum_j\sum_k (x_{ijk} - \bar{x}_{...})^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_R^2$$

avec :

$$S_A^2 = qr\sum_i(\bar{x}_{i..} - \bar{x}_{...})^2$$

$$S_B^2 = pr\sum_j(\bar{x}_{.j.} - \bar{x}_{...})^2$$

$$S_{AB}^2 = r\sum_i\sum_j(\bar{x}_{ij.} - \bar{x}_{.j.} - \bar{x}_{i..} + \bar{x}_{...})^2$$

$$S_R^2 = \sum_i\sum_j\sum_k(x_{ijk} - \bar{x}_{ij.})^2$$

donc que les sommes des carrés s'ajoutent : il y a orthogonalité pour le modèle équilibré.

Remarque : Les modèles orthogonaux sont tels que : $n_{ij} = \dfrac{n_{i.}n_{.j}}{n}$.

Comme en analyse de variance à un facteur, si l'hypothèse H_0 $m_{ij} = 0\ \forall ij$ est vraie, les différentes sommes de carrés suivent à σ^2 près des lois du χ^2 indépendantes. On peut donc tester l'existence des effets principaux, et de l'interaction en comparant S_A^2, S_B^2, S_{AB}^2 à S_R^2. On présente usuellement les résultats sous la forme du tableau 14.4 :

TABLEAU 14.4

Source de variation	Somme de carrés	Degré de liberté (ddl)	Carré moyen	F
A	S_A^2	$p - 1$	$S_A^2/(p-1)$	$\dfrac{S_A^2/p-1}{S_R^2/pq(r-1)}$
B	S_B^2	$q - 1$	$S_B^2/(q-1)$	$\dfrac{S_B^2/q-1}{S_R^2/pq(r-1)}$
Interaction AB	S_{AB}^2	$(p-1)(q-1)$	$\dfrac{S_{AB}^2}{(p-1)(q-1)}$	$\dfrac{S_{AB}^2/(p-1)(q-1)}{S_R^2/pq(r-1)}$
Résiduelle R	S_R^2	$pq(r-1)$	$S_R^2/pq(r-1)$	
Totale	S^2	$pqr - 1$		

14.5.2.3 L'estimation des effets

En posant :

$$\sum_i\alpha_i = \sum_j\beta_j = \sum_i\gamma_{ij} = \sum_j\gamma_{ij} = 0$$

on peut estimer les α_i, β_j, γ_{ij} on trouve alors :

$$\hat{\alpha}_i = \bar{x}_{i..} - \bar{x}_{...}$$
$$\hat{\beta}_j = \bar{x}_{.j.} - \bar{x}_{...}$$
$$\hat{\gamma}_{ij} = \bar{x}_{ij.} - \bar{x}_{i..} + \bar{x}_{...}$$

14.5.2.4 Le cas du plan sans répétition

Le modèle complet avec interaction ne peut être testé et estimé que si et seulement si il y a répétitions car le degré de liberté de S_R^2 est $pq(r-1)$ donc r doit être strictement supérieur à 1.

Si $r = 1$ on doit se contenter du modèle purement additif sans interaction :

$$\mu_{ij} = \alpha_i + \beta_j$$

L'équation d'analyse de variance s'écrit alors :

$$\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = S_A^2 + S_B^2 + S_R^2$$

avec :

$$S_A^2 = q\sum_{i=1}^{p}(\bar{x}_{i.} - \bar{x}_{..})^2$$

$$S_B^2 = p\sum_{j=1}^{q}(\bar{x}_{.j} - \bar{x}_{..})^2$$

$$S_R^2 = \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

On compare donc S_A^2 et S_B^2 au terme d'interaction pris pour terme résiduel ce qui empêche de tester l'interaction.

14.6 TESTS ET PROCÉDURES D'AJUSTEMENT

Ces tests ont pour but de vérifier qu'un échantillon provient ou non d'une variable aléatoire de distribution connue $F_0(x)$.

Soit $F(x)$ la fonction de répartition de la variable échantillonnée, il s'agit donc de tester $H_0 : F(x) = F_0(x)$ contre $H_1 : F(x) \neq F_0(x)$.

Avant de présenter les tests statistiques les plus courants (test du χ^2, de Kolmogorov, de Cramer-Von Mises) il convient de présenter brièvement les procédures empiriques usuelles qui sont une étape indispensable et permettent de s'orienter vers une distribution adaptée aux données.

14.6.1 Les méthodes empiriques

14.6.1.1 La forme de l'histogramme

Celle-ci peut conduire à éliminer certains modèles, en particulier si les propriétés de symétrie ne sont pas vérifiées. Une forme symétrique conduit souvent à poser l'hypothèse de normalité mais il faut se souvenir que la loi de Laplace-Gauss n'est pas la seule à avoir une courbe de densité en cloche : c'est également le cas des lois de Cauchy et de Student entre autres.

Une forme fortement dissymétrique peut suggérer l'usage de lois log-normales, gamma, Weibull ou bêta de type deux qui ont des courbes de densité assez ressemblantes au moins pour certaines valeurs des paramètres.

Le choix entre différentes distributions de forme semblable doit s'effectuer alors en tenant compte du phénomène étudié : ainsi en fiabilité on se limitera aux lois exponentielles ou de Weibull qui ont une justification physique alors que la loi log-normale n'en possède pas dans ce cas.

14.6.1.2 Vérification sommaire de certaines propriétés mathématiques

On vérifiera sur l'échantillon si certaines relations concernant les paramètres d'un modèle sont vraies.

Ainsi pour une loi de Poisson on sait que $E(X) = V(X)$; on s'assurera que sur un échantillon $\bar{x}$ diffère peu de s^{*2}. Une telle constatation est seulement un indice du caractère poissonnien d'une distribution mais n'en est nullement une preuve. On ne peut d'ailleurs jamais prouver la véracité d'un modèle par des moyens statistiques. Un modèle est choisi pour sa commodité et sa faculté de représenter un phénomène.

Pour une variable de Gauss on sait que le coefficient d'aplatissement de cette loi est égal à 3 et que son coefficient d'asymétrie est nul. On vérifiera sur l'échantillon que les coefficients empiriques correspondants s'écartent peu des valeurs théoriques : on dispose pour cela de tables donnant les valeurs critiques de ces coefficients pour différentes tailles d'échantillon (tables A.15 et A.16), voir également plus loin l'abaque pour le test de normalité.

14.6.1.3 Ajustements graphiques

Pour la plupart des lois de probabilité une transformation fonctionnelle simple permet de représenter la courbe de répartition par une droite.

La fonction de répartition empirique d'un échantillon de taille n diffère peu, si n est grand, de la fonction théorique $F(x)$. On vérifiera alors simplement l'adéquation des données au modèle en comparant la fonction de répartition empirique à une droite sur un papier à échelles fonctionnelles.

- **Loi exponentielle**

Si la durée de vie X d'un composant est telle que :

$$P(X > x) = \exp(-\lambda x) \quad \text{on a alors} \quad \ln(1 - F(x)) = -\lambda x$$

Pour un échantillon de taille n on reportera donc pour chaque valeur du temps de fonctionnement x le pourcentage de « survivants » à la date x sur une échelle logarithmique. En pratique on reporte, si les x_i sont ordonnées par valeurs croissantes, les points de coordonnées :

$$x_i \, ; \ln\left(1 - \frac{i-1}{n}\right) \qquad \text{pour} \qquad 1 \leq i \leq n$$

Les points doivent alors être alignés approximativement le long d'une droite dont la pente fournit une estimation graphique de λ.

- Loi de Weibull

Ici $P(X > x) = \exp(-\lambda x^{\beta})$, d'où :

$$\ln(-\ln P(X > x)) = \ln \lambda + \beta \ln x$$

et on reporte les points de coordonnées :

$$\ln x_i \; ; \ln\left(-\ln\left(1 - \frac{i-1}{n}\right)\right)$$

La pente de la droite fournit une estimation graphique de β et son ordonnée à l'origine une estimation de $\ln \lambda$.

- Loi de Laplace-Gauss

Ici la fonction de répartition n'ayant pas d'expression mathématique simple on utilise la propriété $U = \frac{X - m}{\sigma}$ de la manière suivante :

Si les observations x_i proviennent d'une variable normale LG(m ; σ) alors les $u_i = \frac{(x_i - m)}{\sigma}$ constituent un échantillon d'une variable normale centrée-réduite U. Si le nombre des observations est grand, la fonction de répartition empirique (de l'échantillon) doit peu différer de la fonction de répartition théorique telle qu'elle est issue des tables.

Appelons F_i les valeurs de la fonction de répartition empirique $\left(F_i = \frac{\text{effectif} < x_i}{n}\right)$. A ces valeurs empiriques F_i associons les valeurs correspondantes u_i^* de la variable normale centrée réduite obtenues par la table : alors si la distribution est réellement gaussienne et si n est grand, u_i^* doit peu différer de $\frac{(x_i - m)}{\sigma}$ et il doit donc exister une relation linéaire entre u_i^* et x_i (le graphe u_i^*, x_i doit être à peu près une droite coupant l'axe des abscisses en m et de pente $1/\sigma$). Cette droite est appelée la ***droite de Henry***, ou "QQ plot" pour quantile-quantile, en anglais.

Les données ayant été ordonnées par valeurs croissantes, on reportera comme ordonnée de chaque valeur x_i $\frac{i - 3/8}{n + 1/4}$ et non i/n pour des raisons trop compliquées pour être développées ici.

- Exemple

Reprenons les données étudiées au chapitre 5 : les variations du taux de la taxe d'habitation de 100 villes françaises. L'histogramme et le la boîte à moustaches indiquent une répartition plutôt symétrique ; est-elle gaussienne pour autant ?

La droite de Henry montre des écarts importants concernant les queues de distribution ; on peut mettre en doute la normalité de la distribution, mais il ne s'agit pas d'un véritable test où on maitrise les risques d'erreur :

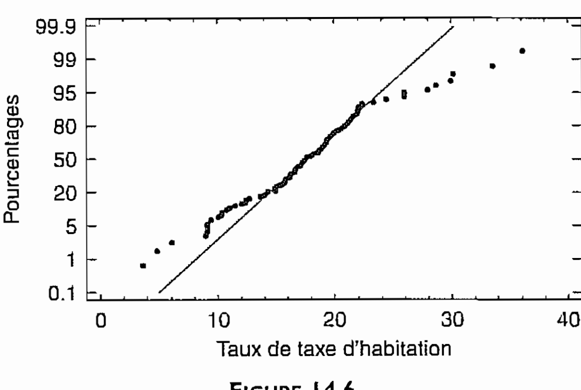


FIGURE 14.6

14.6.2 Les tests statistiques généraux

Nous présenterons ici les deux tests les plus classiques, celui du χ^2 et celui de Kolmogorov-Smirnov, ainsi que celui de Cramer-von Mises.

14.6.2.1 Le test du χ^2

Soit une variable aléatoire X discrète ou discrétisée, c'est-à-dire divisée en k classes de probabilités $p_1, p_2, \ldots, p_k$. Soit un échantillon de cette variable fournissant les effectifs aléatoires $N_1, N_2, \ldots, N_k$ dans chacune de ces classes.

On a $E(N_i) = np_i$.

Considérons alors la statistique D^2 définie comme suit :

$$D^2 = \sum_{i=1}^{i=k} \frac{(N_i - np_i)^2}{np_i}$$

Il est clair que cette statistique est une mesure (on pourrait dire une distance) de l'écart aléatoire entre les effectifs réalisés et les effectifs espérés et intuitivement on sent que D^2 ne peut être trop grand. D^2 dépend du nombre de termes de la somme k mais on remarque que tous ces termes ne sont pas indépendants puisque $\sum_{i=1}^{n} N_i = n$; il suffit d'en connaître en fait $k - 1$.

Donc en fait D^2 dépend de $k - 1$, nombre de degrés de liberté de D^2.

D'après le résultat établi au chapitre 4 à propos de la loi multinomiale on sait que :

THÉORÈME

Si $n \to \infty$, D^2 est asymptotiquement distribué comme une variable de χ^2_{k-1} et ceci quelle que soit la loi de X.

D'où le test du χ^2 : on rejettera H_0 si d^2 constaté $= \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i}$ est trop grand, c'est-à-dire supérieur à une valeur qui n'a qu'une probabilité α d'être dépassée par une variable χ^2.

- **Cas des estimations**

Il arrive bien souvent que seule la forme de la distribution soit spécifiée, Poisson, Laplace-Gauss, mais qu'on ignore certains paramètres que l'on estime sur l'échantillon. Soit « l » le nombre d'estimations indépendantes ainsi réalisées. Le degré de liberté du χ^2 devient alors $k - 1 - l$.

Il convient ici de prendre certaines précautions : les estimations en question doivent être des estimations du maximum de vraisemblances effectuées au moyen des k classes de la distribution, faute de quoi la distribution limite de D^2 n'est plus un χ^2, mais en tout état de cause, comprise entre les valeurs d'un χ^2_{k-1} et d'un χ^2_{k-1-l} : si k est grand, ce phénomène n'est pas trop important, mais si k est petit il peut aboutir à garder inconsidérément H_0 en se fondant sur la distribution de χ^2_{k-1-l}.

- **Effectifs par classes**

La loi de D^2 est asymptotique et l'on admet que $D^2 \sim \chi^2_{k-1}$ si np_i est supérieur à 5 pour toute classe (certains auteurs donnent comme condition 3, ou même 1 pour une seule classe en queue de distribution).

Dans le cas contraire on procédera à des regroupements.

- **Cas des variables continues**

Si on a le choix du découpage en classes, on peut hésiter entre des classes équiprobables et des classes d'égales amplitudes, mais ces dernières doivent être déterminées *a priori*.

Cependant pour des variables continues, le test de Kolmogorov-Smirnov est préférable, s'il n'y a pas d'estimation à effectuer.

Pour des compléments, consulter Kendall et Stuart, volume 2, chapitre 30, *Tests of fit*.

- **Propriétés du test**

On peut démontrer que le test du χ^2 présenté ici est asymptotiquement équivalent au test du rapport des vraisemblances maximales appliqué aux hypothèses :

$$\begin{cases} H_0 : p_i = p_{i0} \ \forall i \\ H_1 : p_i \neq p_{i0} \ \exists i \end{cases}$$

14.6.2.2 Le test d'ajustement de Kolmogorov

Il s'agit d'un test non paramétrique d'ajustement à une distribution entièrement spécifiée de fonction de répartition $F(x)$.

Ce texte repose sur les résultats de Glivenko, Kolmogorov cités en théorie de l'échantillonnage (chapitre 12).

Si F_n^* représente la fonction de répartition empirique d'un n-échantillon d'une variable aléatoire de distribution $F(x)$, on sait que $D_n = \sup|F_n^*(x) - F(x)|$ est asymptotiquement distribué comme suit : $P(\sqrt{n}D_n < y) \to \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2y^2) = K(y)$.

La fonction $K(y)$ a été tabulée et fournit donc un test de :

$$\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$$

La région critique étant définie par $D_n > d(n)$.

Au seuil $\alpha = 0.05$ et si $n > 80$, la région critique est $D_n > \dfrac{1.3581}{\sqrt{n}}$ pour $\alpha = 0.01$ $D_n > \dfrac{1.6276}{\sqrt{n}}$.

Si $n < 80$ on se reportera alors à la table A.14.

14.6.2.3 Le test d'ajustement de Cramer-von Mises

La statistique : $$n\omega_n^2 = \int_{-\infty}^{+\infty} [F_n^*(x) - F(x)]^2 \, \mathrm{d}F(x)$$

est une variable aléatoire dont la distribution indépendante de $F(x)$ sert à tester $H_0 : F(x) = F_0(x)$ contre $H_1 : F(x) \neq F_0(x)$ car $n\omega_n^2$ est une mesure de l'écart existant entre une répartition théorique et une répartition empirique. Sa distribution a été tabulée (voir recueil de tables, table A.13).

On démontre que : $$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_i)\right]^2$$

si les x_i sont les valeurs ordonnées de l'échantillon ($x_1 < x_2, \ldots, < x_n$).

On rejette H_0 si $\dfrac{1}{12n} + \sum_{i=1}^{n}\left[\dfrac{2i-1}{2n} - F_0(x_i)\right]^2$ est supérieur à une valeur que la variable aléatoire $n\omega_n^2$ a une probabilité α de dépasser.

Au seuil $\alpha = 0.05$ on rejette H_0 si $n\omega_n^2 > 0.46136$ pour n grand.

Bien que les lois des statistiques D_n et $n\omega_n^2$ ne soient pas connues, lorsque certains paramètres sont estimés on utilisera avec profit les résultats empiriques (tableau 14.5) obtenus par simulation (Biometrika Tables, volume 2) :

TABLEAU 14.5

Test de normalité	Test d'exponentialité
$H_0 : LG(m, \sigma)$	$H_0 : f(x) = \frac{1}{\theta}\exp\left(\frac{-x}{\theta}\right)$
m est estimé par $\bar{x}$	
σ est estimé par $\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$	θ est estimé par $\bar{x}$
On rejette H_0 :	On rejette H_0 :
– au seuil $\alpha = 0.05$ si $(\sqrt{n} + 0.85/\sqrt{n} - 0.01)D_n > 0.895$	– au seuil $\alpha = 0.05$ si $(D_n - 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n}) > 1.094$
ou $(1 + 0.5/n)n\omega_n^2 > 0.126$	ou $(1 + 0.16/n)n\omega_n^2 > 0.224$
– au seuil $\alpha = 0.01$ si $(\sqrt{n} + 0.85/\sqrt{n} - 0.01)D_n > 1.035$	– au seuil $\alpha = 0.01$ si $(D_n - 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n}) > 1.308$
ou $(1 + 0.5/n)n\omega_n^2 > 0.178$	ou $(1 + 0.16/n)n\omega_n^2 > 0.337$

14.6.3 Exemples d'application en fiabilité et en phénomènes d'attente

14.6.3.1 Test du caractère exponentiel d'une loi de survie

- Expérience classique

On dispose d'un échantillon de n matériels identiques et on note les durées de vie en heures $x_1, x_2, \ldots, x_n$.

Exemple numérique : $n = 5$

$x_1 = 133 \qquad x_2 = 169 \qquad x_3 = 8 \qquad x_4 = 122 \qquad x_5 = 58$

Le paramètre x est estimé par $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 98$, la fonction de répartition estimée est $F(x) = 1 - \exp\left(-\frac{x}{98}\right)$ d'où le tableau :

x_i	8	58	122	133	169
$F(x_i)$	0.079	0.447	0.711	0.743	0.821

La statistique de Kolmogorov vaut :

$$D_n = \sup\left\{\left|F(x_i) - \frac{i}{n}\right| ; \left|F(x_i) - \frac{i-1}{n}\right|\right\}$$

car le maximum est nécessairement atteint en un des points de sauts de la fonction de répartition empirique. On trouve $D_n = 0.311$ soit $\left(D_n - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right) = 0.84$. L'hypothèse H_0 d'une distribution exponentielle peut donc être retenue (avec $\alpha = 0.05$).

La statistique de Cramer-von Mises vaut :

$$n\omega_n^2 = \frac{1}{60} + \sum_{i=1}^{5}\left(\frac{2i-1}{10} - F(x_i)\right)^2 = 0.09133$$

et la quantité $\left(1 + \frac{0.16}{n}\right)n\omega_n^2 = 0.0943$ conduit elle aussi à accepter H_0.

• Expérience de durée limitée avec renouvellement du matériel défaillant

Il est souvent pratiquement impossible de mener à bien l'expérience précédente dès que n est assez élevé car le temps d'étude devient prohibitif. On préfère de beaucoup imposer une durée limite T à l'expérience en renouvelant au besoin au fur et à mesure les appareils tombés en panne de manière à obtenir plus d'informations.

Les instants des pannes obéissent alors, si la durée de vie est exponentielle à un processus de Poisson : en effet si $n = 1$ l'appareil en panne étant remplacé immédiatement, les instants des pannes successives suivent un processus de Poisson car les intervalles entre pannes successives sont indépendants et de loi γ_1 ; pour n appareils, le processus total est une superposition de n processus de Poisson indépendants, ce qui fournit encore un processus de Poisson.

Soit $t_1, t_2, \ldots, t_k$ les instants des pannes pendant T ; d'après le chapitre II, la distribution conditionnelle des dates de panne, sachant k, est une loi uniforme sur $[0, T]$; les instants des pannes $t_1, t_2, \ldots, t_k$ étant ordonnés, les t_i/T forment un échantillon ordonné d'une loi uniforme sur $[0, 1]$ si la durée de vie est exponentielle. Le test du caractère exponentiel de la distribution revient alors à tester l'hypothèse que les t_i/T suivent une loi uniforme sur $[0, 1]$, ce qui peut se faire soit par le test de Kolmogorov, soit par celui de Cramer-von Mises.

Exemples : 100 appareils sont constamment en service et sur une période de 200 heures ; 5 pannes ont été relevées aux instants : $t_1 = 51$, $t_2 = 78$, $t_3 = 110$, $t_4 = 135$, $t_5 = 180$.

- **Test de Kolmogorov**

On cherche le plus grand écart en valeur absolue entre la fonction $F(x) = x$ et les valeurs de la fonction de répartition empirique (fig. 14.7).

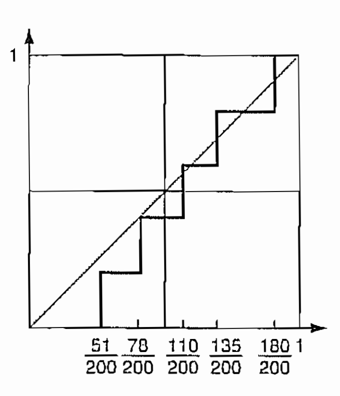


FIGURE 14.7

L'écart est le plus grand pour la première valeur et vaut $D = \frac{51}{200} = 0.255$.

En se reportant à la table de la distribution du test de Kolmogorov on voit qu'on peut accepter l'hypothèse H_0 que la durée de vie obéit à une loi exponentielle pour tout seuil α inférieur à 0.20, puisque à $\alpha = 0.20$, la valeur critique est 0.447.

- **Test de Cramer-von Mises**

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - \frac{t_i}{T} \right]^2$$

puisque pour une loi uniforme sur [0, 1] $F(x) = x$:

$$n\omega_n^2 = \frac{1}{60} + \left(\frac{1}{10} - \frac{51}{200}\right)^2 + \left(\frac{3}{10} - \frac{78}{200}\right)^2 + \left(\frac{5}{10} - \frac{110}{200}\right)^2 + \left(\frac{7}{10} - \frac{135}{200}\right)^2$$
$$+ \left(\frac{9}{19} - \frac{180}{200}\right)^2 \qquad n\omega_n^2 = 0.05192$$

D'après la table A.13, on peut accepter H_0 pour tout seuil α inférieur à 0.85 :

$$P(n\omega_n^2 < 0.447) = 0.95 \; ; \quad P(n\omega_n^2 < 0.056) = 0.15$$

14.6.3.2 Test du caractère poissonnien des arrivées à une file d'attente

Pendant 100 intervalles de 10 min on a compté le nombre X d'ouvriers se présentant à un magasin pour emprunter des outils, ce qui donne le tableau 14.6 (Kaufmann et Faure, *Initiation à la recherche opérationnelle*, Dunod).

On veut vérifier le caractère poissonnien de la loi de X :

On utilisera ici un test du χ^2, car la distribution est discrète (rappelons que les tests de Kolmogorov et de Cramer-Von Mises ne s'appliquent que pour des distributions continues).

TABLEAU 14.6

x_i	n_i	$100p_i$	$\frac{(n_i - 100p_i)^2}{100p_i}$
5	1 ⎫	0.18 ⎫	
6	0 ⎪	0.33 ⎪	
7	1 ⎬ 5	0.74 ⎬ 5.22	0.009
8	2 ⎪	1.45 ⎪	
9	1 ⎭	2.52 ⎭	
10	3	3.93	0.220
11	5	5.58	0.060
12	6	7.26	0.219
13	9	8.72	0.009
14	10	9.73	0.007
15	11	10.12	0.077
16	12	9.87	0.460
17	8	9.07	0.126
18	9	7.86	0.165
19	7	6.46	0.045
20	5	5.04	0.000
21	4	3.75	0.017
22	3 ⎫	2.66 ⎫	
23	1 ⎪	1.80 ⎪	
24	1 ⎬ 6	1.17 ⎬ 7.37	0.255
25	1 ⎪	0.73 ⎪	
≥25	0 ⎭	1.01 ⎭	
			$d^2 = 1.59$

On estime le paramètre λ de la loi de Poisson supposée, par la moyenne empirique qui vaut 15.61. Pour calculer la valeur de D^2 on opère des regroupements aux extrémités pour les classes d'effectifs trop faibles, ce qui laisse 14 classes.

Le paramètre λ ayant été estimé non pas sur les classes résultantes mais sur les valeurs initiales de l'échantillon, la valeur critique pour D^2 est comprise entre celle d'un χ^2_{12} et celle d'un χ^2_{13}.

La valeur du d^2 calculé est bien en deçà de tout seuil de probabilité habituel pour un test : on peut accepter l'hypothèse d'une distribution poissonnienne.

Remarque : un esprit soupçonneux trouverait peut être cette valeur de d^2 trop faible, puisque d'après les tables il y a 995 chances sur 1 000 pour que χ^2_{12} soit supérieur à 3. L'ajustement est-il trop beau pour être vrai ? Nous laisserons le lecteur juge . . ..

14.6.4 Tests de normalité

L'ajustement d'une distribution normale à des données réelles justifie un traitement particulier en raison de l'importance de la loi normale.

Il est tout à fait déconseillé d'utiliser le test du khi-deux en raison de son manque de puissance et du caractère subjectif du découpage en classes.

On peut utiliser les variantes des tests de Kolmogorov et Cramer-Von Mises indiquées précédemment, mais ces tests *omnibus* n'utilisent pas de propriétés spécifiques de la loi de Gauss et sont moins puissants que les suivants qui sont d'ailleurs recommandés par la norme AFNOR NF X-06-050.

Le plus simple à utiliser est le test conjoint d'asymétrie et d'aplatissement qui se présente sous forme d'abaque (*cf.* annexe).

Il suffit de vérifier si le point dont l'abscisse est la valeur absolue du coefficient d'asymétrie (skewness) et l'ordonnée le coefficient d'aplatissement (kurtosis) se situe à l'intérieur, donc vers la gauche de la courbe correspondant à la taille d'échantillon.

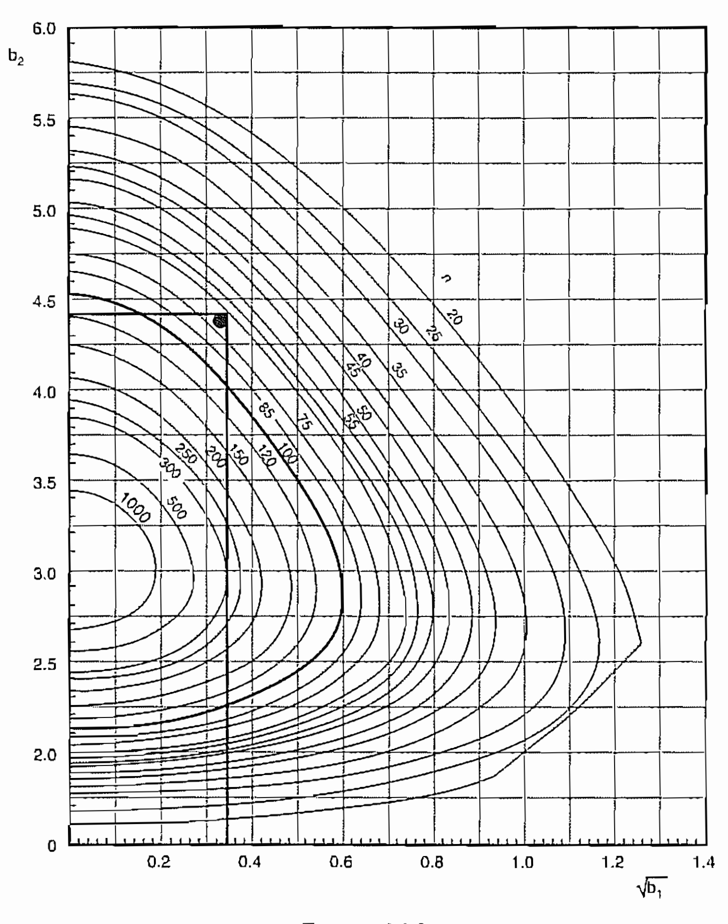


FIGURE 14.8

Pour l'exemple des 100 valeurs de la taxe d'habitation on trouve un coefficient d'asymétrie de 0.34 et un coefficient d'aplatissement de 4.47 ce qui conduit à rejeter l'hypothèse de normalité.

Le test le plus recommandé est celui de Shapiro-Wilk basé sur la comparaison de l'écart-type à une combinaison linéaire $\Sigma a_i w_i$ des étendues successives $w_1 = x_n - x_1$; $w_2 = x_{n-1} - x_2$ etc.

Les coefficients de la combinaison dépendent de n et sont données dans des tables, ainsi que la région critique. Les calculs sont programmés dans la plupart des logiciels statistiques.

Dans l'exemple précédent les statistiques de Kolmogorov et de Cramer-Von Mises valent respectivement 0.105 et 0.195 et conduisent au rejet de l'hypothèse de normalité.

Rappelons pour conclure que d'une part ne pas rejeter une hypothèse ne prouve pas sa véracité et que plusieurs distributions peuvent s'ajuster aux mêmes données, et d'autre part que lorsque que la taille de l'échantillon augmente il devient de plus en plus difficile d'accepter un modèle standard de distribution : en effet les lois de probabilité d'usage courant sont des modèles simplificateurs à peu de paramètres et ne peuvent pas toujours rendre compte de la complexité d'un grand ensemble de données.

14.7 QUELQUES LIMITES DES TESTS

Au terme de ce long chapitre il convient de faire les remarques suivantes. Les tests sont un outil essentiel du processus de décision en faveur ou en défaveur d'une hypothèse scientifique. Cela a pu conduire à des excès : un résultat ne pouvait être publié dans certaines revues que si un test l'avait déclaré statistiquement significatif au risque 5 %.

Un point essentiel concerne la taille des échantillons : l'inférence statistique classique a été développée pour traiter des « petits » échantillons de l'ordre de quelques dizaines ou centaines d'observations au plus. En présence de très grandes bases de données le paradoxe est que tout devient significatif : par exemple, pour un million d'individus, l'hypothèse d'indépendance entre deux variables sera rejetée au risque 5 % si le coefficient de corrélation linéaire est supérieur en valeur absolue à 0.002, ce qui est sans intérêt pratique. On peut considérer que l'hypothèse nulle a été mal choisie, mais le problème persiste : l'hypothèse nulle devant être fixée avant la collecte, ou en tous cas avant l'analyse des données, on aboutira à son rejet dès qu'elle est trop précise car tout écart même minime devient significatif.

Le problème se pose dans les mêmes termes pour les tests d'ajustement à des modèles : si les données sont des données réelles et non simulées, on aura tendance à rejeter le modèle. Il ne faut pas s'en étonner puisqu'un modèle est une simplification de la réalité : comment imaginer que l'on puisse représenter des millions d'observations avec seulement 2 ou 3 paramètres ? Ce n'est pas d'ailleurs pour cela qu'il faut nécessairement abandonner le modèle, qui peut avoir d'autres vertus . . . L'analyse des grandes bases de données amène ainsi à repenser la notion de test et conduit à des recherches nouvelles sur la validation (voir chapitre 19).

Un autre problème se pose quand on effectue un très grand nombre de tests sur les mêmes données, par exemple en génétique pour savoir si certains caractères sont présents. Il s'agit d'un cas semblable aux comparaisons multiples (voir paragraphe 14.5.1.4) mais de grande ampleur. Le risque de rejeter à tort une des hypothèses nulles croît rapidement avec le nombre de tests. Il faut alors recourir à la théorie du contrôle du taux de fausses découvertes (Benjamini et Hochberg, 1995).

15 Méthodes de Monte-Carlo et de rééchantillonnage (Jack-knife, bootstrap)

Dans de nombreux cas, il n'est pas possible d'obtenir des expressions exactes pour les distributions de statistiques de test ou d'estimateurs, car les calculs sont trop complexes. Les méthodes de simulation et de rééchantillonnage qui ont pu se développer avec les progrès de l'informatique permettent de substituer à une étude théorique impossible, une démarche expérimentale où les lois exactes sont approchées par des répartitions empiriques. La simulation aléatoire (dite de Monte-Carlo par référence aux jeux de hasard) consiste à reproduire avec un ordinateur de nombreux échantillons issus de lois connues et à effectuer pour chacun de ces échantillons les calculs nécessaires, qui sont ensuite synthétisés.

15.1 GÉNÉRATION DE VARIABLES ALÉATOIRES

À la base des méthodes de Monte-Carlo se trouve la nécessité de simuler des échantillons artificiels de variables aléatoires. Toutes les méthodes reposent sur la génération de variables uniformes.

15.1.1 Génération de variables uniformes sur [0 ; 1]

Bien qu'il existe des procédés physiques de réalisation de variables uniformes (roue de loterie par exemple) ils ne sont guère compatibles avec l'informatique et la nécessité de disposer très rapidement de grands échantillons. On recourt donc à des algorithmes de génération de valeurs comprises entre 0 et 1 : un algorithme étant par nature déterministe, on parle alors de nombres pseudo-aléatoires. Un bon algorithme doit pouvoir réaliser des suites très grandes de nombres qui ont en apparence toutes les propriétés d'un n-échantillon de variables indépendantes et identiquement distribuées.

Il est donc important de tester la qualité d'un générateur : on utilise pour cela des tests classiques d'ajustement et d'indépendance.

Les méthodes les plus employées sont basées sur des suites récurrentes (qui fournissent donc nécessairement des suites périodiques). La méthode multiplicative congruentielle de Lehmer est la plus connue : $r_{i+1} = ar_i$ modulo m, c'est-à-dire que r_{i+1} est le reste de ar_i divisé par m.

En pratique on prend m le plus grand possible afin d'avoir la période la plus grande possible.

On peut montrer que si a est de forme $8t \pm 3$ et si r_0 est un nombre entier positif impair quelconque la période de la suite engendrée est $m/4$.

Les nombres $\dfrac{r_i}{m-1}$ compris entre 0 et 1 sont alors considérés comme pseudoaléatoires, c'est-à-dire comme un échantillon de la loi uniforme sur [0, 1].

Sur ordinateur on choisira généralement $m = 2^{p-1}$ où p est le nombre de bits d'un mot machine (le premier bit est inutilisable car réservé au signe). De plus la division par 2^{p-1} est aisée à faire sur ordinateur car elle correspond à une troncature.

Un choix classique est $a = 7^5 = 16\,807$, ou $a = 2^{16} + 3 = 65\,539$ avec $m = 2^{31} - 1$

15.1.2 Méthodes générales de tirage d'un échantillon artificiel de *n* valeurs d'une variable aléatoire *X* continue

15.1.2.1 Inversion de la fonction de répartition

La méthode suivante s'applique lorsque F^{-1} a une forme analytique simple.

Soit $F(x)$ la fonction de répartition de X. La variable $Y = F(X)$ est uniformément distribuée sur [0, 1].

En effet :

$$g(y) = \frac{f[F^{-1}(y)]}{F'[F^{-1}(y)]} = 1$$

Donc si l'on tire n nombres au hasard uniformément répartis entre 0 et 1 : $r_1, r_2, \ldots, r_n$ l'échantillon cherché $(x_1, x_2, \ldots, x_n)$ sera déterminé par $x_i = F^{-1}(r_i)$; cette méthode est dite « de l'anamorphose » (fig. 15.1).

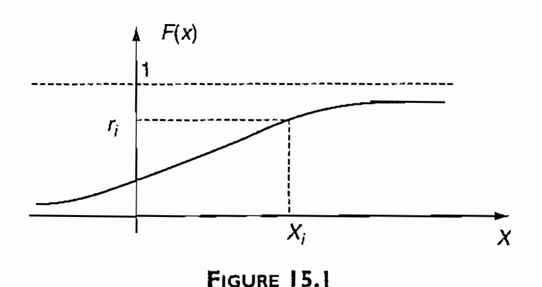


FIGURE 15.1

On dit que l'on a simulé la variable X.

15.1.2.2 Méthode du rejet de von Neumann

Cette méthode est applicable lorsque la densité de X est à support borné et reste finie. On supposera que $0 \leq X \leq 1$.

Soit m un majorant de $f(x)$. On tire un nombre U uniformément réparti entre 0 et 1 et ensuite un nombre V uniformément réparti entre 0 et m (fig. 15.2).

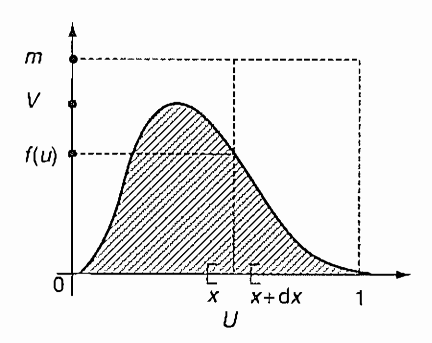


FIGURE 15.2

Si $V < f(U)$ on conserve U qui est une réalisation de X et on pose $u = x$ sinon on rejette U et on recommence.

En effet la probabilité conditionnelle qu'une valeur U soit gardée sachant que $U = x$ vaut $P(V < f(x)) = f(x)/m$, la probabilité inconditionnelle est $1/m$ (rapport de la surface sous la courbe de densité à la surface du rectangle), donc d'après la formule de Bayes :

$$P(x < U < x + dx / U \text{ est gardée}) = \frac{\frac{f(x)}{m}\,dx}{1/m} = f(x)\,dx$$

Cette méthode est recommandée pour simuler les lois bêta de type I dont on peut déduire la loi bêta de type II par la transformation $Y = X/1 - X$.

La méthode du rejet peut conduire dans certains cas à rejeter un trop grand nombre de valeurs.

Une amélioration notable de la méthode du rejet consiste à utiliser une autre fonction de densité g facilement simulable, telle que $cg(x) \geq f(x)$. On génére alors un couple $(y\ ;\ u)$ de réalisations indépendantes de Y de densité g et de u uniforme. Si $u < \dfrac{f(y)}{cg(y)}$, y est accepté comme réalisation de X de densité $f(x)$. Sinon on rejette la valeur y et on recommence. Si X est à support borné, on prendra par exemple pour Y une loi triangulaire.

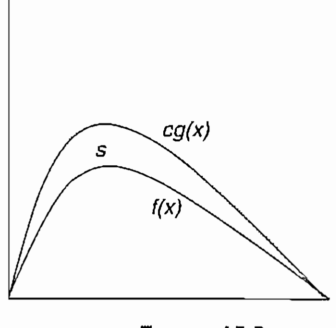


FIGURE 15.3

Si s est la surface comprise entre les deux courbes, on montre que le taux de valeurs acceptées vaut $\frac{1}{s+1}$.

Lorsque la distribution à simuler est trop complexe, ou que sa densité n'est connue qu'à un facteur multiplicatif près comme dans le cas de distributions *a posteriori* en statistique bayésienne, on utilise actuellement des méthodes dites MCMC (Monte-Carlo Markov Chains) qui consistent à simuler une chaîne de Markov qui converge vers la distribution à simuler. Les MCMC dépassent le cadre de cet ouvrage et nous renvoyons le lecteur intéressé à l'ouvrage de C. Robert (2001).

15.1.3 Méthodes spécifiques

15.1.3.1 Variable de Bernoulli X de paramètre p

On tire un nombre au hasard entre 0 et 1 : r.

Si $r < p$, $X = 1$; si $r > p$, $X = 0$.

À partir de la variable de Bernoulli on peut simuler une variable binomiale $\mathcal{B}(n\,;p)$ en utilisant le fait qu'une binomiale est une somme de n variables de Bernoulli indépendantes.

15.1.3.2 Loi γ_p avec p entier

La fonction de répartition d'une variable X suivant une loi γ_1 est $F(x) = 1 - \exp(-x)$. La méthode de l'anamorphose nous donne, si r est un nombre aléatoire uniformément distribué entre 0 et 1 : $r = 1 - \exp(-x)$ soit $x = -\ln(1 - r)$. Comme $1 - r$ est aussi uniformément distribué sur [0, 1] il suffit pour simuler X de faire :

$$x = -\ln r$$

Une variable X suivant une loi γ_p est une somme de p variables γ_1 indépendantes, d'où la formule de simulation : $x = -\ln r_1 - \ln r_2 \cdots - \ln r_p$ si p est entier soit :

$$x = -\ln\left(\prod_{i=1}^{n} r_i\right)$$

15.1.3.3 Loi de Poisson $\mathcal{P}(\lambda)$

La méthode consiste à simuler un processus de Poisson de cadence 1 sur une période égale à λ puisque le nombre d'événements survenant sur cette période suit une loi de Poisson $\mathcal{P}(\lambda)$ (fig. 15.3).

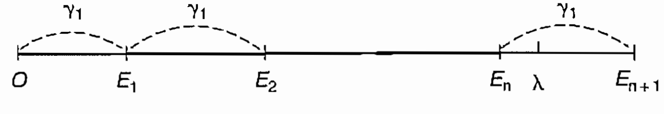


FIGURE 15.4

Les intervalles successifs OE_1 ; $E_1E_2, \ldots E_nE_{n+1}$ suivent indépendamment des lois γ_1. On engendre comme en 15.1.3.2 des variables γ_1 et on ajoute leurs valeurs jusqu'à

dépasser λ ; la réalisation n de la variable de Poisson $\mathcal{P}(\lambda)$ est alors le plus grand entier n tel que : $\sum_{i=1}^{i=n} - \ln r_i < \lambda$ ou ce qui est équivalent mais plus économique du point de vue calcul :

$$\prod_{i=1}^{n} r_i > \exp(-\lambda)$$

15.1.3.4 Variable de Laplace-Gauss

- La méthode suivante repose sur le théorème central-limite

$\dfrac{\bar{X}\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} LG(0 ; 1)$. Ce théorème étant valable en particulier pour des variables uniformes, la somme de n variables uniformes est donc approximativement une loi de Laplace-Gauss d'espérance $n/2$ et de variance $n/12$ car la loi continue uniforme sur [0, 1] a pour espérance $1/2$ et pour variance $1/12$.

En pratique ce résultat est acquis dès que $n = 12$ d'où la méthode :

Pour obtenir une réalisation d'une variable LG(6 ; 1) ajouter 12 nombres au hasard tirés entre 0 et 1.

Soit $r_1, r_2, \ldots, r_{12}$ ces nombres et soit X une variable LG(m ; σ) ; on a alors :

$$x = m + \sigma\left(\sum_{i=1}^{12} r_i - 6\right)$$

- Méthode de Box et Müller

Cette méthode exacte découle du théorème suivant :

Si U et V sont deux variables uniformes sur [0, 1] indépendantes, alors X et Y définies par :

$$X = (-2 \ln U)^{1/2} \cos 2\pi V$$
$$Y = (-2 \ln U)^{1/2} \sin 2\pi V$$

sont deux variables normales centrées-réduites indépendantes.

En effet en notant $\rho^2 = X^2 + Y^2$ et $\theta = \text{Arc tg}\,\dfrac{Y}{X}$, ρ^2 et θ suivent indépendamment des lois χ_2^2 et uniforme sur [0, 2π]. L'algorithme de Box-Muller revient à simuler l'angle θ par $2\pi V$ et le rayon ρ par $(-2 \ln U)^{1/2}$ puisque $\chi_2^2/2$ suit une loi exponentielle (voir chapitre 4, paragr. 4.3.1).

D'où pour deux nombres aléatoires r_1 et r_2, deux réalisations de la loi LG(0 ; 1) indépendantes.

- Méthode polaire de Marsaglia

C'est une variante de la précédente utilisant une technique de rejet qui évite le calcul des sinus et cosinus.

On engendre deux nombres aléatoires r_1 et r_2 puis $u_1 = 2r_1 - 1$ et $u_2 = 2r_2 - 1$ (u_1 et u_2 sont uniformément répartis sur l'intervalle $[-1 ; +1]$). On rejette u_1 et u_2 si $u_1^2 + u_2^2 > 1$ afin de garder un couple uniformément réparti dans le cercle de rayon unité (fig. 15.5).

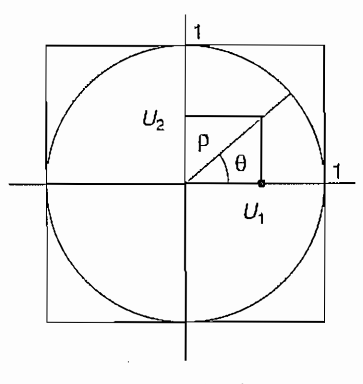


FIGURE 15.5

On montre alors que $\rho^2 = U_1^2 + U_2^2$ est une variable uniforme sur [0, 1] indépendante du couple $\left(\frac{U_1}{\rho}, \frac{U_2}{\rho}\right)$ où $\frac{U_1}{\rho}$ et $\frac{U^2}{\rho}$ sont le cosinus et le sinus de l'angle aléatoire θ uniformément réparti entre 0 et 2π ; d'où la formule :

$$x_1 = \sqrt{\frac{-2 \ln \rho^2}{\rho^2}}\, u_1 \qquad \text{et} \qquad x_2 = \sqrt{\frac{-2 \ln \rho^2}{\rho^2}}\, u_2$$

La simulation d'un vecteur aléatoire gaussien dont les composantes ne sont pas indépendantes peut s'effectuer en recourant à une ACP : si l'on connaît la matrice de variance covariance Σ, on en cherche les vecteurs propres qui vont fournir des combinaisons linéaires gaussiennes et indépendantes que l'on simule aisément. On peut également utiliser la transformation de Mahalanobis. Il suffit ensuite de faire faire la transformation inverse (voir chapitre 4).

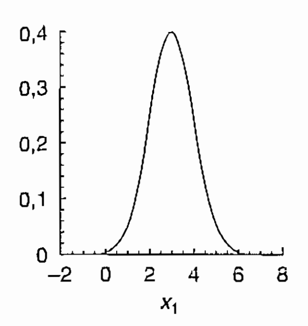
15.2 APPLICATIONS

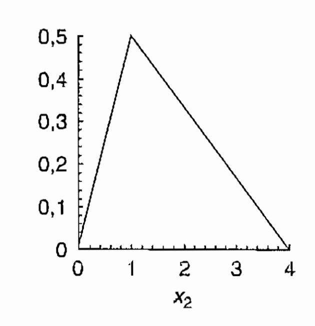
15.2.1 Simulation de fonctions de variables aléatoires

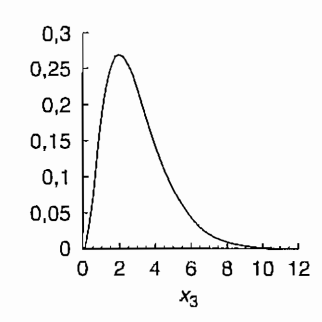
Soit Y une variable s'exprimant comme une fonction $f(X_1, X_2, \ldots X_p)$ de variables de lois connues. Il sera en général difficile de trouver la loi (densité ou fonction de répartition) de Y même dans des cas simples. Si les X_i sont indépendantes, il est facile d'obtenir un échantillon artificiel de Y : il suffit de générer indépendamment une valeur de chaque variable, de calculer f et de recommencer. On peut ainsi résoudre le problème du calcul d'incertitudes en physique ou chimie où l'on connaît l'incertitude sur chaque variable sous la forme Δx qui

n'est en fait que deux fois l'écart-type de X si l'on se réfère à la loi normale. L'incertitude Δy sera estimée par la moitié de l'écart-type de Y.

■ **Exemple :** $Y = \dfrac{X_1 X_2}{X_3}$ où X_1 suit une loi normale N(3 ; 1), X_2 une loi triangulaire et X_3 une loi gamma de paramètre 3. X_1, X_2, X_3 sont des variables indépendantes.

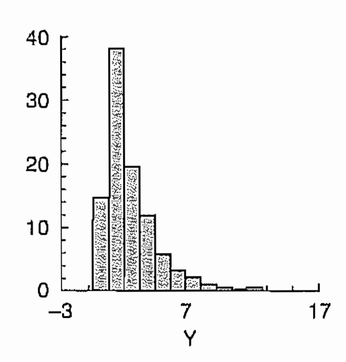






En simulant 1000 valeurs de Y on trouve :

Moyenne = 2,4915
Médiane = 1,66771
Variance = 8,71593
Écart-type = 2,95227
Minimum = –0,0406886
Maximum = 48,2368
Étendue = 48,2775
Asymétrie = 5,79394
Aplatissement = 65,2381



Avec 1000 réalisations, les résultats sont suffisamment précis :

Intervalle de confiance à 95, 0 % pour la moyenne : [2,30852 2,67448]

Intervalle de confiance à 95, 0 % pour l'écart-type : [2,82831 3,08768] ■

15.2.2 Calcul d'une intégrale par la méthode de Monte Carlo

Toute intégrale peut se ramener par un changement de variable à une intégrale entre 0 et 1.

Or $I = \displaystyle\int_0^1 g(t)\,\mathrm{d}t$ est l'espérance de $g(U)$ où U est une variable uniforme sur [0, 1].

A partir d'un échantillon de la loi uniforme U, on estimera I par $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(u_i)$ moyenne des valeurs de la variable $g(U)$.

On a : $$E(\hat{I}) = I \quad \text{et} \quad V(\hat{I}) = \frac{1}{n}V(g(U)) = \frac{1}{2n}\iint [g(u) - g(v)]^2 \,\mathrm{d}u\,\mathrm{d}v$$

Le procédé peut être amélioré en remarquant que :

$$I = \int_0^1 \frac{g(t)}{p(t)}p(t)\,\mathrm{d}t$$

où $p(t)$ est la densité d'une variable T définie sur [0, 1] ; un choix judicieux de $p(t)$ appelée « fonction d'importance » permet de diminuer considérablement la variance de l'estimation.

En effet $I = E\left[\frac{g(T)}{p(T)}\right]$ d'où : $$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\frac{g(t_i)}{p(t_i)}$$

où les points t_i ne sont plus choisis uniformément entre 0 et 1 mais selon la loi de densité p.

On a alors : $$V(\hat{I}) = \frac{1}{2n}\iint\left[\left(\frac{g(t)}{p(t)}\right) - \left(\frac{g(u)}{p(u)}\right)\right]^2 \mathrm{d}t\,\mathrm{d}u$$

La variance est alors nulle si p est proportionnel à g (ce qui suppose I connu . . .).

En pratique on prendra une fonction d'importance dont l'allure est voisine de celle de g.

Ce type de calcul est surtout utile pour l'évaluation numérique d'intégrales multiples.

15.2.3 Distributions d'échantillonnage de statistiques complexes

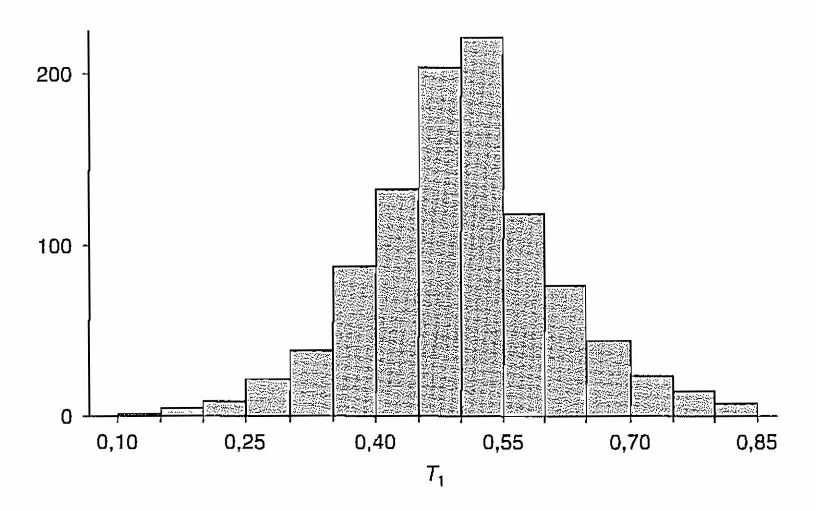
La simulation permet également de trouver la distribution approchée de statistiques complexes, et même de statistiques simples quand la population a une distribution peu maniable.

Il suffit de répéter N fois la simulation d'un n-échantillon de X pour obtenir N valeurs de la statistique d'intérêt T : si N est assez grand, on aura une bonne précision.

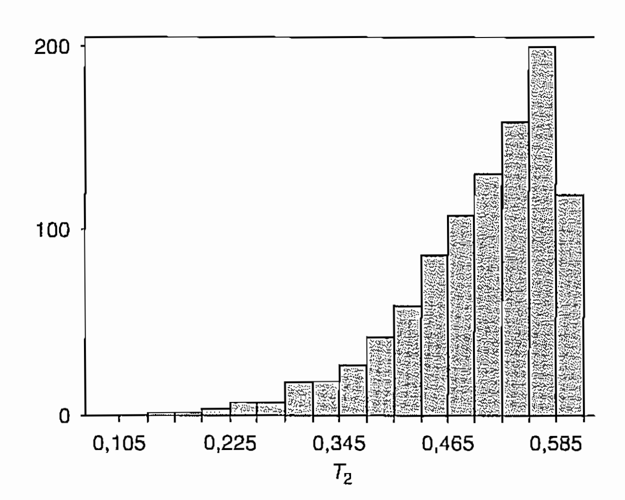
Exemple : Si U est une variable aléatoire uniforme sur [0 ; θ] le milieu de l'étendue d'un n-échantillon est un estimateur de $\theta/2$ que l'on notera T. Quelle est la distribution et la variance de T pour $n = 5$?

En prenant $\theta = 1$, on simule $N = 1000$ échantillons de 5 valeurs d'une loi uniforme, on calcule à chaque fois T, d'où sa distribution empirique.

Une simulation donne $\bar{t} = 0{,}5003752 \quad s = 0{,}1106459$



Remarque : ce n'est pas l'estimateur sans biais de variance minimale qui est $\frac{n+1}{2n}\sup(X_1 ; \ldots ; X_n)$ et dont la distribution est la suivante :



15.2.4 Données manquantes et imputation multiple

Il est fréquent d'avoir des valeurs manquantes dans des tableaux de données (données omises, refus de réponse, erreurs etc.). Avant de savoir comment traiter le problème, il faut s'interroger sur le mécanisme qui a produit une valeur manquante pour une variable Y. Pour simplifier, nous nous placerons dans le cas où une seule variable numérique présente une valeur manquante. Le mécanisme est dit « ***non-ignorable*** » si la probabilité que Y soit manquant dépend de la vraie valeur de Y (exemple : la probabilité de ne pas donner son revenu est d'autant plus grande que le revenu est élevé). Des modèles spécifiques sont alors nécessaires pour prendre en compte ce mécanisme.

Dans le cas contraire on dira que la donnée est manquante aléatoirement (« ***missing at random*** »). Deux options principales s'offrent au praticien :

– ignorer la donnée manquante en supprimant l'individu de l'analyse, mais on voit vite que s'il faut supprimer tous les individus dans ce cas, on risque d'appauvrir fortement l'échantillon ;
– remplacer la valeur manquante par une valeur plausible : c'est l'***imputation***.

Il existe de nombreuses méthodes d'imputation :

– remplacer la valeur manquante par la moyenne des valeurs non-manquantes (mais on ne tient pas compte des autres variables) ;
– effectuer une régression multiple où Y est expliquée par les autres variables sur les données complètes.

Ces méthodes dites d'imputation simple souffrent d'un défaut majeur : elles sont déterministes en ce sens que deux individus qui ont les mêmes valeurs des autres variables auront la même valeur imputée de Y, ce qui n'est pas réaliste et conduit à une diminution artificielle de la variance. Il vaut mieux tirer au hasard une réalisation de Y, considérée comme une variable aléatoire, dans la loi conditionnelle de $Y/X_1, X_2, \ldots X_p$, d'où l'utilisation des techniques de simulation.

La solution la plus élaborée rendue possible par les moyens de calcul actuels est l'***imputation multiple*** : on effectue plusieurs tirages, ce qui conduit à plusieurs tableaux de données que l'on analyse séparément. Les résultats sont ensuite regroupés pour étudier la variabilité attribuable aux données manquantes.

Le problème est en réalité assez complexe et nécessite une approche bayésienne : si l'on utilise un modèle de régression pour estimer la valeur manquante $Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \varepsilon$, il ne suffit pas de tirer des valeurs dans la distribution du résidu ε, mais il faut tenir compte du fait que les coefficients β_j du modèle sont estimés, donc aléatoires. On doit donc tirer aussi des valeurs des β_j dans leur distribution *a posteriori* qui elle même dépend des valeurs manquantes.

Nous renvoyons le lecteur intéressé à l'excellent petit livre de Paul D. Allison (2001).

15.3 MÉTHODES DE RÉÉCHANTILLONNAGE

Les méthodes de simulation exposées précédemment permettent d'obtenir des distributions d'échantillonnage d'estimateurs dans le cas classique où l'on dispose d'un modèle paramétrique $f(x\ ;\ \theta)$. En l'absence de modèle réaliste, ce qui est souvent le cas en pratique, comment simuler ? En d'autres termes comment tirer des réalisations d'une distribution inconnue ? Ce problème paradoxal se résout en tirant dans une distribution proche de la distribution inconnue et la meilleure en l'absence d'information, n'est autre que la distribution empirique. C'est le principe des méthodes de rééchantillonnage où on va tirer au hasard des observations dans l'échantillon dont on dispose. Dans le bootstrap on effectue des tirages avec remise, alors que le jack-knife procède par tirages sans remise.

15.3.1 Le bootstrap

Soit une variable X de loi F inconnue ; on dispose d'un échantillon $(x_1, x_2, \ldots, x_n)$ et on veut étudier par exemple la distribution d'un estimateur T d'un certain paramètre θ, calculer sa variance, en donner un intervalle de confiance.

L'idée de cette méthode due à B. Efron repose sur le principe élémentaire suivant :

Si n est grand F_n^* est proche de F, on aura donc une bonne approximation de la loi de T en utilisant F_n^* à la place de F.

On est donc amené à tirer des échantillons de n valeurs dans la loi F_n^* ce qui revient à rééchantillonner dans l'échantillon $x_1, x_2, \ldots, x_n$; autrement dit à effectuer des tirages avec remise de n valeurs parmi les n valeurs observées : les valeurs observées $x_1, x_2, \ldots, x_n$ sont donc répétées selon les réalisations d'un vecteur multinomial $K_1, K_2, \ldots, K_n$ d'effectif n et de probabilités p_i égales à $1/n$.

Lorsque n n'est pas très élevé on peut énumérer tous les échantillons possibles équiprobables (il y en a n^n) sinon on se contente d'en tirer un nombre B suffisament grand à l'aide d'une technique de tirage dans une population finie.

Si le nombre de réplications B tend vers l'infini, la moyenne de toutes les estimations bootstrap converge vers l'estimateur du maximum de vraisemblance empirique (c'est-a-dire utilisant la loi F_n^*) et permet ainsi d'estimer sa variance. En pratique on se contentera de quelques centaines de tirages au plus.

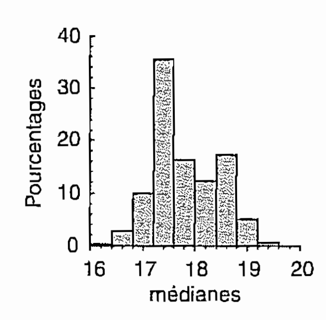
Le calcul d'intervalles de confiance peut se faire par :

- La méthode des percentiles qui consiste simplement à repérer les quantiles souhaités dans la distribution des B valeurs. C'est la méthode la plus simple.
- L'approximation normale : on calcule la moyenne et l'écart-type des B réalisations et on utilise un intervalle à $\pm 1{,}96$ écart-types. Il faut bien sur vérifier la normalité approximative de la distribution des B valeurs.
- D'autres méthodes plus complexes nécessitant une estimation de la variance pour chaque échantillon répliqué (voir Davison & Hinkley, 1997).

Exemple : bien que ce ne soit pas un échantillon aléatoire, reprenons les données du chapitre 5 sur les valeurs du taux de taxe d'habitation de 100 communes françaises. On souhaite obtenir un intervalle de confiance pour la médiane qui vaut 17,625 et on effectue pour cela 1000 retirages.

On trouve la distribution suivante :

Moyenne = 17.7872
Médiane = 17.625
Écart-type = 0.630658
Minimum = 15.87
Maximum = 19.39



L'intervalle de confiance des percentiles est [16,70 18,92] en prenant respectivement la 25e et la 975e valeur ordonnée.

L'intervalle avec approximation normale [16,55 19,02].

Le bootstrap est donc une méthode très générale qui permet de répondre à des problèmes jusque là quasi impossibles à résoudre comme l'étude de la variabilité de résultats d'analyses factorielles (valeurs propres, vecteurs propres etc.) ou l'estimation de variance dans des sondages complexes. Il faut cependant être conscient que si la taille n de l'échantillon de départ est faible, il y aura en général sous-estimation de la variabilité : les intervalles de confiance auront tendance à être trop petits (couverture insuffisante). En effet le rééchantillonnage ne permet pas par définition d'engendrer des valeurs autres que celles déjà observées, ce qui peut être gênant pour des variables numériques, mais l'est moins pour des variables qualitatives où en général, toutes les modalités sont observées, au moins marginalement. Cela étant, le bootstrap est une méthode d'étude de la variabilité intrinsèque à un échantillon.

15.3.2 Le Jack-knife

Cette technique a été proposée par Quenouille pour diminuer le biais d'un estimateur et reprise par Tukey ; elle est moins performante que le bootstrap.

15.3.2.1 Définition

Soit T un estimateur calculé sur un échantillon de taille n.

On note T_{-i} l'estimateur calculé sur le $(n - 1)$ échantillon obtenu en enlevant l'observation i et on appelle pseudo-valeur T_i^* :

$$\boxed{T_i^* = nT - (n-1)T_{-i}}$$

L'estimateur *Jack-knife* est alors la moyenne des pseudo-valeurs :

$$\boxed{T_J = \frac{1}{n}\sum_{i=1}^{n} T_i^*}$$

ce qui donne $T_J = T - (n-1)\frac{1}{n}\sum_{i=1}^{n}(T_i - T)$.

La variance de l'estimateur *Jack-knife* est alors donnée par :

$$\boxed{S_J^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(T_i^* - T_J)^2}{n-1}}$$

15.3.2.2 Réduction du biais

Supposons que $E(T) = \theta + \frac{a}{n}$, alors $E(T_J) = \theta$.

En effet :
$$\begin{aligned} E(T_J) &= E(T) - (n-1)(E(T_{-i}) - E(T)) \\ &= \theta + \frac{a}{n} - (n-1)\left[\theta + \frac{a}{n-1} - \theta - \frac{a}{n}\right] \\ &= \theta + \frac{a}{n} - a + \frac{n-1}{n}a = \theta \end{aligned}$$

À titre d'exercice on peut vérifier que la méthode du *Jackknife* appliquée à la variance S^2 donne l'estimateur S^{*2}, et que appliquée à $\bar{x}$ on retrouve $\bar{x}$. Le calcul du *Jackknife* est surtout utile pour des statistiques biaisées dont le biais est très difficile à calculer (coefficient de corrélation par exemple).

15.3.2.3 Intervalle de confiance

J. Tukey a émis la conjecture suivante :

$$\frac{T_J - \theta}{S_J} = T_{n-1}$$

qui permettrait d'obtenir des intervalles de confiance indépendamment de toute hypothèse sur la loi de X et en se servant uniquement de l'information apportée par les données. Cependant cette conjecture est manifestement fausse dans certains cas : la médiane en particulier car les T_{-i} ne peuvent prendre que deux valeurs différentes (si n est pair).

Il vaut mieux prendre comme degré de liberté le nombre de pseudo-valeurs réellement distinctes diminué d'une unité, ce qui conduit à des résultats souvent acceptables.

16 La régression simple

Considérons un couple de variables aléatoires numériques (X, Y). Si X et Y ne sont pas indépendantes, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y : elle la diminue en général, car la distribution conditionnelle de Y, sachant que $X = x$, a une variance qui est en moyenne inférieure à la variance de Y : $E[V(Y/X)] \leq V(Y)$ d'après le théorème de la variance totale (il est toutefois possible que $V(Y/X = x) > V(Y)$ pour certaines valeurs de X).

Lorsque l'on peut admettre que le phénomène aléatoire représenté par X peut servir à prédire celui représenté par Y (causalité, concomitance, etc.), on est conduit à rechercher une formule de prévision de Y par X du type $\hat{Y} = f(X)$, sans biais $E[Y - \hat{Y}] = 0$, ainsi qu'à évaluer l'ordre de grandeur de l'erreur de prévision que l'on mesure par la variance de $\varepsilon = Y - \hat{Y}$. On cherchera bien sûr à minimiser cette variance.

Nous étudierons le cas théorique en recherchant la formule de prévision idéale (au sens des moindres carrés), plus spécialement si cette formule est linéaire avec un écart-type conditionnel constant $\sigma(\varepsilon/X = x) = \sigma$ (homoscédasticité), puis le cas usuel où les variables ne sont connues qu'à travers les valeurs d'un échantillon.

X sera dit variable explicative ou prédicteur ;
Y sera dit variable expliquée ou critère.

Certaines propriétés seront seulement énoncées, le lecteur étant renvoyé aux démonstrations faites dans le chapitre sur la régression multiple.

16.1 LE MODÈLE THÉORIQUE DE LA RÉGRESSION SIMPLE

16.1.1 L'approximation conditionnelle

Étant donné deux variables aléatoires Y et X, la recherche d'une fonction f telle que $f(X)$ soit aussi proche que possible de Y en moyenne quadratique a déjà été abordée au chapitre 3, paragraphe 3.3.2.

On sait que $f(X) = E(Y/X)$ réalise le minimum de $E[(Y - f(X))^2]$ car $E(Y/X)$ est la projection orthogonale de Y sur l'espace L_X^2 des variables du type $f(X)$ (fig. 16.1), espace contenant Δ droite des constantes.

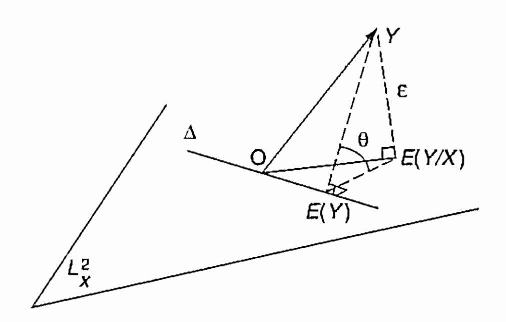


FIGURE 16.1

La qualité de l'approximation de Y par $E(Y/X)$ est mesurée par le rapport de corrélation :

$$\eta^2_{Y/X} = \frac{V(E(Y/X))}{V(Y)} = \frac{\text{Variance expliquée}}{\text{Variance totale}} = \cos^2\theta$$

La fonction qui, à une valeur x de X, associe $E(Y/X = x)$ s'appelle fonction de régression de Y en X, son graphe est la courbe de régression de Y en X.

On peut alors poser $Y = E(Y/X) + \varepsilon$, où ε est un résidu aléatoire pas toujours négligeable.

ε a pour propriété d'être d'espérance nulle : $E(\varepsilon) = 0$ car $E(Y) = E(E(Y/X))$.

De plus, ε est non corrélé linéairement avec X et avec $E(Y/X)$, car ε est orthogonal à L^2_X.

La variance de ε ou variance résiduelle est alors $V(\varepsilon) = (1 - \eta^2_{Y/X})\, V(Y)$.

16.1.2 Cas où la régression est linéaire

Ce cas, le plus important dans la pratique, est celui où $E(Y/X) = \alpha + \beta X$. (Ceci se produit en particulier si X et Y suivent une loi normale à deux dimensions). On a donc :

$$\boxed{Y = \alpha + \beta X + \varepsilon}$$

En prenant l'espérance des deux membres de la relation $E(Y/X) = \alpha + \beta X$, il vient :

$$E(Y) = \alpha + \beta E(X)$$

La droite de régression passe donc par le point de coordonnées $(E(X), E(Y))$. On a :

$$Y - E(Y) = \beta(X - E(X)) + \varepsilon$$

en multipliant par $X - E(X)$ de chaque côté et en prenant l'espérance :

$$E[(Y - E(Y))(X - E(X))] = \beta E[(X - E(X))^2] + E[\varepsilon(X - E(X))]$$

soit $\text{cov}\,(X, Y) = \beta V(X) + \text{cov}\,(\varepsilon, X)$ car $E(\varepsilon) = 0$. Mais, comme ε est non corrélé avec X, il reste :

$$\beta = \frac{\text{cov}(X, Y)}{V(X)} = \rho\,\frac{\sigma_Y}{\sigma_X}$$

L'équation de la droite de régression est donc :

$$E(Y/X) - E(Y) = \frac{\text{cov}(X, Y)}{V(X)}(X - E(X))$$

d'où :

$$\boxed{Y = E(Y) + \rho \frac{\sigma_Y}{\sigma_X}(X - E(X)) + \varepsilon}$$

Comme ε est non corrélé avec X, on peut écrire, en prenant la variance des deux membres :

$$V(Y) = \rho^2 \frac{\sigma_Y^2}{\sigma_X^2} V(X) + V(\varepsilon)$$

$$V(Y) = \rho^2 V(Y) + V(\varepsilon)$$

Comme on a vu que $V(\varepsilon) = (1 - \eta^2_{Y/X})V(Y)$, on retrouve le fait que si la régression est linéaire :

$$\rho^2 = \eta^2_{Y/X}$$

Rappelons que dans ce cas, il est inutile de chercher une transformation de X, autre que linéaire puisque $E(Y/X)$ est la meilleure approximation possible de Y par $f(X)$.

16.2 AJUSTEMENT SUR DES DONNÉES

On dispose de n couples (x_i, y_i) ; $i = 1, \ldots, n$ constituant un n-échantillon d'***observations indépendantes*** de (X, Y). On suppose vraie l'hypothèse :

$$E(Y/X) = \alpha + \beta X$$

Le problème est donc d'estimer α, β ainsi que la variance σ^2 du résidu ε.

La méthode qui va être développée s'applique encore si la variable X n'est pas aléatoire, mais contrôlée par l'expérimentateur (c'est le cas par exemple quand on mesure Y différence de potentiel aux bornes d'une résistance pour différentes valeurs de l'intensité du courant : l'intensité n'est pas aléatoire, mais Y l'est, par suite des erreurs de mesure entre autres), ou imposée par la nature des choses (Y est une grandeur mesurée à différentes dates, $x_1, \ldots, x_n$; X est donc le temps). Il suffit alors de supposer que pour chaque observation, on a $y_i = \alpha + \beta x_i + \varepsilon_i$ où les ε_i sont des réalisations ***indépendantes*** d'une variable ε d'espérance nulle et de ***variance constante*** σ^2, quel que soit x_i.

On parle alors de ***modèle linéaire*** plutôt que de régression linéaire.

C'est parce que les propriétés de la méthode des moindres carrés ne dépendent que des lois conditionnelles à X fixé que l'on peut traiter indifféremment la régression linéaire et le modèle linéaire par les mêmes techniques. On prendra garde cependant de ne parler de corrélation entre Y et X que lorsque X est aléatoire.

De nombreux modèles non linéaires se ramènent facilement au modèle linéaire par des transformations simples.

Ainsi le modèle $y \simeq \alpha x^\beta$, très utilisé en économétrie (élasticité constante de y par rapport à x ; β coefficient d'élasticité), devient un modèle linéaire en passant aux logarithmes : $y' = \ln y$, $x' = \ln x$ et alors $y' \simeq \ln \alpha + \beta x'$.

Il en va de même pour le cas du modèle à croissance exponentielle : $y \simeq \alpha \exp(\beta x)$; il suffit de poser $y' = \ln y$ pour avoir $y' \simeq \ln \alpha + \beta x$.

Le modèle logistique souvent posé pour rendre compte des variations d'un taux de réponse y (compris entre 0 et 1) en fonction d'une « excitation » x : $y \simeq \dfrac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$ se linéarise en posant $y' = \ln \dfrac{y}{1-y}$. On a alors $y' = \alpha + \beta x$.

Cependant le modèle $y \simeq \alpha + \exp(\beta x)$ n'est pas linéarisable, tandis que le modèle $y \simeq \alpha + \beta x + \gamma x^2$ est linéaire, mais est à deux variables explicatives si on pose $x^2 = z$ et $y \simeq \alpha + \beta x + \gamma z$ (voir régression multiple).

16.2.1 Estimation de α, β, σ^2 par la méthode des moindres carrés

La méthode des moindres carrés due à Gauss reprend sur l'échantillon la propriété que $E(Y/X) = \alpha + \beta X$ est la meilleure approximation de Y par X en moyenne quadratique. On cherche donc à ajuster au nuage des points (x_i, y_i) une droite d'équation $y^* = a + bx$ de telle sorte que $\sum_{i=1}^{n} (y_i - y_i^*)^2$ soit minimal (fig. 16.2).

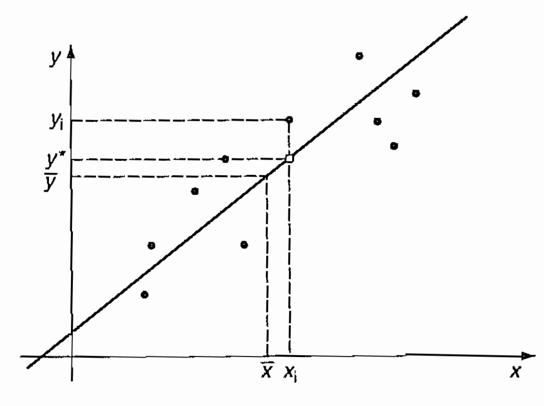


FIGURE 16.2

On étudie ensuite les propriétés de a et b en tant qu'estimations de α et β ainsi que l'estimation $\hat{\sigma}^2$ de σ^2 que l'on en déduit.

La méthode élémentaire de détermination de a et b est la suivante :

$$\sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = F(a, b)$$

Ce minimum est atteint pour $\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$, ce qui donne les deux équations :

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \Leftrightarrow \overline{Y} = a + b\overline{x}$$

$$\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0$$

dont la solution est :

$$\boxed{b = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = r\frac{s_y}{s_x} = \frac{\text{cov}(x, y)}{s_x^2}}$$

d'où :

$$\boxed{y^* = \overline{y} + r\frac{s_y}{s_x}(x - \overline{x})}$$

La droite des moindres carrés passe donc par le centre de gravité du nuage et sa pente est l'analogue empirique de la pente de la droite de régression $\rho\frac{\sigma_y}{\sigma_x}$.

Puisque les y_i et, dans le cas de la régression, les x_i, sont des réalisations de variables aléatoires, il ne faut pas perdre de vue que $\overline{x}, \overline{y}, r, s_x, s_y, a, b$, sont des réalisations de variables aléatoires.

THÉORÈME I

a, b et y^* sont des estimations sans biais de α, β et de $E(Y/X = x) = \alpha + \beta x$.

b est une réalisation de la variable aléatoire B :

$$B = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Pour montrer que $E(B) = \beta$, nous allons montrer en fait que $E^{(x_i)}(B) = \beta$ où $E^{(x_i)}(B)$ désigne l'espérance conditionnelle de B connaissant les valeurs $X_i = x_i$ des variables X_i. Comme l'espérance de l'espérance conditionnelle est l'espérance de B on aura $E(B) = \beta$ *a fortiori* :

$$E^{(x_i)}(B) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})E^{(x_i)}(Y_i - \overline{Y})}{\sum(x_i - \overline{x})^2}$$

Or $E^{(x_i)}(Y_i) = \alpha + \beta x_i$ par hypothèse de régression linéaire, et aussi alors :

$$E^{(x_i)}(\overline{Y}) = \alpha + \beta\overline{x}$$

Donc : $E^{(x_i)}(Y_i - \overline{Y}) = \beta(x_i - \overline{x})$; $E^{(x_i)}(B) = \beta \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

on a donc : $\boxed{E^{(x_i)}(B) = \beta}$ d'oú $\boxed{E(B) = \beta}$

Comme $a = \overline{y} - b\overline{x}$, a est une réalisation de $A = \overline{Y} - B\overline{X}$, et, par le même procédé :

$$E^{(x_i)}(A) = E^{(x_i)}(\overline{Y}) - \overline{x}E^{(x_i)}(B)$$
$$= \alpha + \beta\overline{x} - \overline{x}\beta$$

$\boxed{E^{(x_i)}(A) = \alpha}$ donc $\boxed{E(A) = \alpha}$

Puisque $E(Y/X = x) = \alpha + \beta x$, $y^* = a + bx$ est une estimation sans biais de $\alpha + \beta x$.

On peut montrer de plus que B n'est pas corrélé avec $\overline{Y}$: on a tout d'abord la simplification suivante :

$$b = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x})y_i}{\sum_i (x_i - \overline{x})^2}$$

car : $\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i (x_i - \overline{x})y_i - \overline{y}\sum_i (x_i - \overline{x})$ et $\sum_i (x_i - \overline{x}) = 0$

La covariance conditionnelle de B et $\overline{Y}$ à x_i fixés est donc :

$$\text{cov}(B\,;\overline{Y}) = \text{cov}\left(\frac{\sum_i (x_i - \overline{x})Y_i}{\sum_i (x_i - \overline{x})^2}\,;\overline{Y}\right) = \frac{1}{\sum_i (x_i - \overline{x})^2}\sum_i (x_i - \overline{x})\,\text{cov}(Y_i\,;\overline{Y})$$

Or $\text{cov}(Y_i\,;\overline{Y}) = \text{cov}\left(Y_i\,;\frac{1}{n}\sum_j Y_j\right) = \frac{\sigma^2}{n}$, car Y_i et Y_j sont indépendants si $i \neq j$; il vient :

$$\text{cov}(B\,;\overline{Y}) = \frac{\sigma^2}{n\sum_i (x_i - \overline{x})^2}\sum_i (x_i - \overline{x}) = 0$$

car $\sum_i (x_i - \overline{x}) = 0$. B et $\overline{Y}$ sont non corrélés conditionnellement aux x_i, ils le sont donc marginalement.

Cependant, le fait d'être sans biais n'est qu'une qualité mineure pour des estimateurs. Le théorème suivant (pour une démonstration, voir le chapitre sur la régression multiple) prouve la qualité des estimations obtenues, ceci sans référence à aucune loi de probabilité.

THÉORÈME 2 (GAUSS-MARKOV)

A et B sont parmi les estimateurs sans biais de α et β fonction linéaire des Y_i, ceux de variance minimale.

Montrons que la variance conditionnelle de B est :

$$V^{(x_i)}(B) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

En effet : $$V^{(x_i)}(B) = V^{(x_i)}\left(\frac{\sum_i (x_i - \bar{x})(Y_i)}{\sum_i (x_i - \bar{x})^2}\right) = \frac{\sum_i (x_i - \bar{x})^2 V^{(x_i)}(Y_i)}{\left(\sum_i (x_i - \bar{x})^2\right)^2} = \frac{\sum_i (x_i - \bar{x})^2 \sigma^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2}$$

puisque à x_i fixé, $Y_i = \alpha + \beta x_i + \varepsilon$. Donc $V(Y_i/X_i = x_i) = V(\varepsilon) = \sigma^2$.

Comme $A = \overline{Y} - B\overline{X}$ on a $V(A) = V(\overline{Y}) + \bar{x}^2 V(B)$ à x_i fixés, d'où :

$$V^{x_i}(A) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)$$

Pour exprimer $\sigma^2 = V(\varepsilon)$, il est naturel de penser à utiliser la variance des résidus $e_i = y_i - y_i^*$, c'est-à-dire la quantité que l'on a minimisée : $\sum_i (y_i - y_i^*)^2$.

On montre alors (*cf.* régression multiple) le théorème :

THÉORÈME 3

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - y_i^*)^2}{n - 2}$$ *est une estimation sans biais de* σ^2.

16.2.2 Propriétés des écarts résiduels

Soit $e_i = y_i - y_i^*$ l'écart résiduel.

THÉORÈME

Les e_i sont de moyenne nulle.

Démonstration

Comme $y_i^* = \bar{y} + b(x_i - \bar{x})$, on a $\sum_i e_i = \sum_i (y_i - y_i^*) = \sum_i (y_i - \bar{y}) - b\sum_i (x_i - \bar{x})$ donc $\sum_i e_i = 0$, ce qui prouve que les e_i ***ne sont pas*** des réalisations ***indépendantes*** d'une variable aléatoire.

La variance empirique des e_i est donc égale à $\frac{1}{n}\sum_i e_i^2$ et est notée $s^2_{y/x}$ et est appelée variance résiduelle.

On a alors le résultat suivant :

$$\boxed{s^2_{y/x} = (1 - r^2)s^2_y}$$

En effet :

$$s^2_{y/x} = \frac{1}{n}\sum_i (y_i - y_i^*)^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2 + \frac{b^2}{n}\sum_i (x_i - \bar{x})^2 - \frac{2b}{n}\sum_i (y_i - \bar{y})(x_i - \bar{x})$$

$$s^2_{y/x} = s^2_y + b^2 s^2_x - 2b\,\text{cov}(x, y)$$

$$= s^2_y + r^2 s^2_y - 2r\frac{s_y}{s_x} r s_y s_x = s^2_y + r^2 s^2_y - 2r^2 s^2_y = (1 - r^2)s^2_y$$

16.2.3 Cas où le résidu ε suit une loi normale

Tous les résultats établis précédemment supposaient uniquement $E(Y/X) = \alpha + \beta X$.

Si on admet maintenant que ε suit une loi LG(0 ; σ), on a tout d'abord :

a) $Y/X = x \in \text{LG}(\alpha + \beta x\,;\,\sigma)$.

b) B, A, Y^* suivent, les x_i fixés, des lois de Laplace-Gauss car ils sont des combinaisons linéaires de lois de Laplace-Gauss :

$$B \in \text{LG}\left(\beta\,;\,\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}\right)$$

$$A \in \text{LG}\left(\alpha\,;\,\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}\right)$$

$$Y^* \in \text{LG}\left(\alpha + \beta x\,;\,\sigma\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}\right)$$

c) A, B, $\hat{\sigma}^2$ sont les estimateurs de variance minimale de α, β, σ^2.

d) $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} = \dfrac{\sum_i (y_i - y_i^*)^2}{\sigma^2} = \dfrac{ns^2_{y/x}}{\sigma^2}$ est une réalisation d'une variable χ^2_{n-2} indépendante de $\bar{Y}$, de B et de A.

Remarque : A et B ne sont pas indépendants.

Les points *c*) et *d*) seront démontrés dans le chapitre concernant la régression multiple.

L'usage des lois de A et B suppose σ connu, ce qui n'est pas vrai en général.

Puisque $\dfrac{(B-\beta)\sqrt{\sum_i (x_i - \bar{x})^2}}{\sigma} \in LG(0,1)$ et $\dfrac{nS^2_{y/x}}{\sigma^2} \in \chi^2_{n-2}$ sont indépendantes on a :

$$\frac{(B-\beta)s_x}{s_{y/x}}\sqrt{n-2} \text{ suit un } T_{n-2}$$

ce qui permet de donner des intervalles de confiance pour β.

La relation précédente s'exprime usuellement par :

$$\boxed{\frac{(B-\beta)}{\hat{\sigma}_b} = \sqrt{n}\,\frac{(B-\beta)s_x}{\hat{\sigma}} = T_{n-2}}$$

On trouve de même : $$\boxed{\frac{(A-\alpha)}{\hat{\sigma}_a} = \frac{(A-\alpha)}{\hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}} = T_{n-2}}$$

Application : Loi de r quand $\rho = 0$, cas de la régression entre variables aléatoires.

Quand $\rho = 0$, on a $\beta = 0$ (hypothèse dite de non-régression).

En remplaçant B par $R\dfrac{S_y}{S_x}$ et $S_{y/x} = S_y\sqrt{1-R^2}$ il vient facilement :

$$\boxed{\frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \text{ suit un } T_{n-2}}$$

16.3 TESTS DANS LE MODÈLE LINÉAIRE

16.3.1 Analyse de variance de la régression

Effectuons la décomposition classique :

$$y_i - \bar{y} = y_i - y_i^* + y_i^* - \bar{y}$$

où ε est supposé $LG(0 ; \sigma)$.

On voit aisément que $\sum_i (y_i^* - \bar{y})(y_i - y_i^*) = 0$.

Donc : $$\boxed{\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{Somme des carrés totale}} = \underbrace{\sum_i (y_i - y_i^*)^2}_{\text{Somme des carrés résiduelle}} + \underbrace{\sum_i (y_i^* - \bar{y})^2}_{\text{Somme des carrés expliquée}}}$$

On sait que :
$$\frac{\sum_i (Y_i - Y_i^*)^2}{\sigma^2} = \chi^2_{n-2}$$

Si l'hypothèse $H_0 : \beta = 0$ (hypothèse de non-régression linéaire) est vraie et dans ce cas seulement :

$$\frac{\sum_i (Y_i - \overline{Y})^2}{\sigma^2} = \chi^2_{n-1}$$

$$\sum_i \left(\frac{Y_i^* - \overline{Y}}{\sigma}\right)^2 = \sum_i \frac{B^2(X_i - \overline{X})^2}{\sigma^2}$$

Puisque $\beta = 0$, on en déduit alors que $\sum_i \frac{(Y_i^* - \overline{Y})^2}{\sigma^2}$ suit un χ^2_1 car on sait que $\frac{(B - \beta)^2 \sum_i (X_i - \overline{X})^2}{\sigma^2}$ suit un χ^2_1 comme carré d'une variable LG(0 ; 1).

Le théorème de Cochran s'applique et $\sum_i (Y_i - Y_i^*)^2$ et $\sum_i (Y_i^* - \overline{Y})^2$ sont donc indépendants et alors :

$$\boxed{\frac{\sum_i (Y_i^* - \overline{Y})^2}{\sum_i (Y_i^* - Y_i)^2}(n - 2) \text{ suit un } F(1\,;\, n - 2) \quad \text{si } \beta = 0}$$

Le test du caractère significatif de la régression est alors immédiat. Ce test est d'ailleurs identique à celui du coefficient de corrélation linéaire :

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

qui, lui, n'a de sens que si X et Y sont aléatoires.

En effet, le rapport précédent n'est autre que $\frac{R^2}{1 - R^2}(n - 2)$ car :

$$\frac{1}{n}\sum_i (Y_i^* - Y_i)^2 = S^2_{Y/X} = (1 - R^2)S^2_Y$$

Et comme l'on sait que $(T_{n-2})^2 = F(1\,;\, n - 2)$ le test précédent est donc équivalent au test portant sur R.

16.3.2 Test d'une équation de régression spécifiée

$$\begin{cases} H_0 : \alpha = \alpha_0 & \text{et} & \beta = \beta_0 \\ H_1 : \alpha \neq \alpha_0 & \text{ou} & \beta \neq \beta_0 \end{cases}$$

Il s'agit de tester simultanément les deux coefficients de régression. Comme A et B ne sont pas indépendants, il serait incorrect de tester successivement α puis β.

Nous donnerons ici uniquement le résultat, qui est un cas particulier de celui obtenu en régression multiple. Un tel test est souvent utile pour savoir si une droite des moindres carrés diffère significativement de la première bissectrice.

Si H_0 est vraie, la quantité $\frac{1}{2\hat{\sigma}^2}\left[n(a-\alpha_0)^2 + 2n\bar{x}(a-\alpha_0)(b-\beta_0) + (b-\beta_0)^2\sum_i x_i^2\right]$ est une réalisation d'une variable $F(2\,;n-2)$. On rejettera H_0 si la quantité trouvée est trop grande.

16.3.3 Test de linéarité de la régression

Ce test, qui semble fondamental, a pour but de savoir si l'hypothèse $E(Y/X) = \alpha + \beta X$ est fondée. Il devrait donc précéder toute étude de régression linéaire. En fait, ce test nécessite d'avoir des observations répétées de Y pour chaque valeur de X, ce qui est souvent difficile sauf dans le cas d'une expérimentation où X est un facteur contrôlé. En effet, on cherche à savoir si la courbe des moyennes conditionnelles est une droite, en d'autres termes si les $\bar{y}_j$, moyennes des n_j observations de Y lorsque $X = x_j$, sont à peu près liées linéairement aux x_j.

Pour cela, on compare le coefficient de corrélation linéaire r^2 au rapport de corrélation empirique : $e^2 = \frac{\frac{1}{n}\sum_j n_j(\bar{y}_j - \bar{y})^2}{s_y^2}$ car, dans l'hypothèse de régression linéaire $\eta^2_{Y/X} = \rho^2$.

On montre alors que si l'hypothèse $H_0 : \eta^2_{Y/X} = \rho^2$ ou $E(Y/X) = \alpha + \beta X$ est vraie alors :

$$\frac{(e^2 - r^2)/k-2}{(1-e^2)/n-k} = F(k-2\,;n-k)$$

où k est le nombre de valeurs distinctes de X. On rejettera H_0 si le rapport est trop grand.

Dans ces conditions, on pourra aussi tester les hypothèses $H_0 : \eta^2_{Y/X} = 0$ contre $H_1 : \eta^2_{Y/X} \neq 0$ afin de savoir si une formule de régression autre que linéaire peut être essayée.

En effet, si $\eta^2_{Y/X} = 0$ est vraie, on sait que $\frac{e^2/k-1}{(1-e^2)/n-k} = F(k-1\,;n-k)$.

16.3.4 Contrôle des hypothèses du modèle linéaire

Les propriétés de la méthode des moindres carrés dépendent essentiellement du fait que le résidu ε a une variance constante quel que soit x, et qu'il n'y a pas d'autocorrélation entre les diverses réalisations de ε.

Il convient donc toujours de s'assurer de la validité de ces deux hypothèses, ce que l'on fait usuellement en étudiant de manière empirique (des tests rigoureux sont délicats à établir) les valeurs des écarts résiduels $e_1, e_2, \ldots, e_n$ qui ne doivent pas laisser apparaître de tendance quand on les confronte graphiquement aux x_i par exemple, ou de dépendance en étudiant la liaison e_i, e_{i+1}. Dans le cas contraire, les estimateurs $b, a, \hat{\sigma}$ ne sont plus de variance minimale (ils restent toutefois sans biais). On se reportera au chapitre suivant pour l'étude détaillée des résidus.

- **Le test de Durbin-Watson**

Ce test est couramment utilisé en économétrie pour s'assurer de la non corrélation des résidus. On suppose ici que les observations sont ordonnées par le temps et on teste l'hypothèse H_0 : « non corrélation des ε_i » contre H_1 : « ε_i processus auto-régressif d'ordre 1 » c'est-à-dire $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$ avec $\rho > 0$ (le cas $\rho < 0$ est en général sans intérêt).

On prend pour statistique de test :

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

qui doit être voisin de 2 si H_0 est vraie.

On vérifie sans peine que $0 \leq d \leq 4$. ρ peut être estimé par $1 - d/2$. Les valeurs critiques de d ont été tabulées (voir annexe table A.17).

- **Un cas simple d'hétéroscédasticité**

Il est fréquent d'avoir $V(\varepsilon/X = x) = \sigma^2 x^2$: l'écart-type du résidu croît linéairement avec le prédicteur.

Les estimateurs des moindres carrés sont sans biais mais ne sont plus de variance minimale. En écrivant la vraisemblance des y_i on a :

$$L(y_1, y_2, \ldots, y_n) = \frac{1}{(2\pi)^{n/2}\sigma^n \prod_{i=1}^{n} x_i} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{y_i - \alpha - \beta x_i}{x_i}\right)^2\right)$$

Les estimateurs du maximum de vraisemblance sont alors les estimateurs de variance minimale et on voit sans peine que le problème est équivalent à une régression usuelle sur des données transformées.

En posant $y_i' = \frac{y_i}{x_i}$, $x_i' = \frac{1}{x_i}$ et $\varepsilon_i' = \frac{\varepsilon_i}{x_i}$ on a en effet : $y_i' = \beta + \alpha x_i' + \frac{\varepsilon_i}{x_i} = \beta + \alpha x_i' + \varepsilon_i'$ avec $V(\varepsilon_i') = \sigma^2$. Il suffit donc d'ajuster une droite au nuage $\left(\frac{y_i}{x_i} ; \frac{1}{x_i}\right)$.

La constante du modèle transformé est la pente de la droite de régression du modèle originel et *vice-versa*.

On obtiendra alors, bien sûr, une analyse de variance de la régression moins flatteuse mais des estimations plus précises des coefficients de régression.

16.4 APPLICATIONS

16.4.1 Exemple (tableau 16.1)

Les données suivantes, communiquées par M. Tenenhaus, professeur à HEC, concernent un échantillon de 24 offres de vente d'appartements situés dans le 5[e] et le 6[e] arrondissements de Paris, en 1975.

Tableau 16.1

Y Prix en milliers de Francs	130	280	800	268	790	500	320	250
X Surface en mètres carrés	28	50	196	55	190	110	60	48
Prix	378	250	350	300	155	245	200	325
Surface	90	35	86	65	32	52	40	70
Prix	85	78	375	200	270	295	85	495
Surface	28	30	105	52	80	60	20	100

La forme du nuage de points autorise un ajustement linéaire (fig. 16.3). On pose donc le modèle $Y = \alpha + \beta X + \varepsilon$ et on supposera $\varepsilon \in LG(0 : \sigma)$.

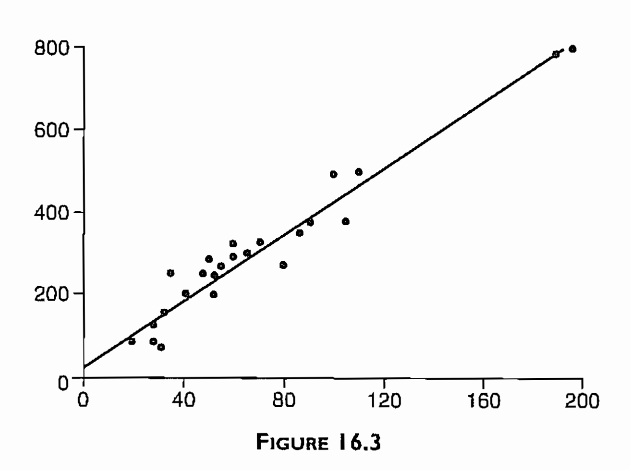


Figure 16.3

Des calculs élémentaires conduisent aux 5 résultats suivants, qui contiennent toute l'information utile :

$$\bar{x} = 70{,}0833 \text{ m}^2 \qquad \bar{y} = 309{,}333\ 10^3 \text{ F} \qquad s_x = 44{,}6915 \text{ m}^2$$

$$s_y = 182{,}9505\ 10^3 \text{ F} \qquad r = 0{,}9733$$

On en déduit tout d'abord les estimations a et b de α et β : $a = 30{,}0921$ et $b = 3{,}9844$.

L'équation de la droite d'ajustement est donc $y^* = 3{,}9844x + 30{,}0921$.

Notons ici que les estimateurs des moindres carrés sont invariants par changement d'échelle des variables au sens suivant :

Si Y est multiplié par une constante k (passage du franc à l'euro par exemple), la pente b et l'ordonnée à l'origine a sont multipliées par la même constante k.

Si X est multiplié par une constante k (par exemple surface exprimée en pieds carrés au lieu de m²), la pente est divisée par k, l'ordonnée à l'origine ne change pas.

Dans tous les cas, le coefficient de corrélation ne change pas, pas plus que les statistiques de test.

La variance résiduelle $s^2_{y/x}$ s'obtient directement par la formule $s^2_{y/x} = (1 - r^2)s^2_y$, soit :

$$s^2_{y/x} = 1762{,}1816 \qquad \text{d'où} \qquad s_{y/x} = 41{,}98$$

On en déduit : $\hat{\sigma}^2 = \dfrac{n}{n-2} s^2_{y/x} = 1922{,}38 \qquad \text{d'où} \qquad \hat{\sigma} = 43{,}84.$

Les estimations des variances de A et B sont donc :

$$\hat{\sigma}^2_a = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{n s^2_x}\right) = (16{,}6455)^2 = 277{,}0724$$

$$\hat{\sigma}^2_b = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{n s^2_x} = (0{,}2003)^2 = 0{,}0401$$

L'hypothèse de normalité de ε permet de donner des intervalles de confiance pour ces diverses estimations ; ainsi, pour σ^2, $\dfrac{n s^2_{y/x}}{\sigma^2}$ est une réalisation d'une variable χ^2_{n-2} ; la table de la distribution de χ^2 à 22 degrés de liberté fournit les bornes 11 et 36,8 pour un intervalle de probabilité à risques symétriques de niveau 0.95 (fig. 16.4).

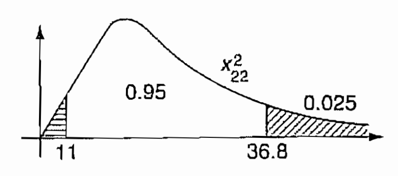


FIGURE 16.4

L'intervalle de confiance est donc donné par :

$$\frac{24 s^2_{y/x}}{36.8} < \sigma^2 < \frac{24 s^2_{y/x}}{11}$$
$$1149{,}25 < \sigma^2 < 3844{,}76$$
$$33{,}90 < \sigma < 62{,}01$$

soit avec $1 - \alpha = 0{,}95$.

Le test de signification de la régression peut être effectué par l'analyse de variance présentée dans le tableau 16.2 :

TABLEAU 16.2

Source de variation	Somme des carrés	Degré de liberté	Carré moyen
Expliquée par la régression	761 009	1	761 009
Résiduelle	42 292	22	1 922,4
Totale	803 301	23	

La valeur f constatée $\dfrac{761\,009}{1\,922{,}4} = 396$ est évidemment très significative.

On pourrait aussi, ce qui est strictement équivalent, tester $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$ en comparant $b/\hat{\sigma}_b$ à une variable de Student à $22 = n - 2$ degrés de liberté :

$$t = \frac{b}{\hat{\sigma}_b} = 19{,}9$$

ce qui excède tout seuil usuel. On peut donc accepter H_0. On aura remarqué que $19{,}9 = (396)^{1/2}$.

Les deux tests précédents sont aussi équivalents au test du coefficient de corrélation linéaire $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$.

16.4.2 Prévision d'une valeur ultérieure

Supposons maintenant que l'on désire prévoir à l'aide du modèle la valeur de Y pour une valeur non observée x_0 de X. La prévision naturelle est $Y_0^* = a + bx_0$.

Afin d'encadrer cette valeur, cherchons ce que l'on appelle un intervalle de prévision.

On a vu que Y_0^* est distribué selon une loi :

$$LG\left(\alpha + \beta x_0 \,;\, \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}\right)$$

On sait d'autre part que la loi de $Y/X = x_0$ (en abrégé Y_0) est une loi $LG(\alpha + \beta x_0 \,;\, \sigma)$ par hypothèse du modèle de régression linéaire. Y_0 et Y_0^* sont deux variables indépendantes, car Y_0 ne dépend que de la valeur future x_0 tandis que Y_0^* ne dépend que des valeurs déjà observées $(x_1, x_2, \ldots, x_n)$ si l'on suppose les réalisations de ε indépendantes.

$Y_0 - Y_0^*$ suit alors une loi $LG\left(0 \,;\, \sigma\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}\right)$ et donc :

$$\frac{Y_0 - Y_0^*}{\hat{\sigma}\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}} \text{ suit une loi de Student } T_{n-2}$$

En remplaçant x_0 par sa valeur et Y_0^* par $a + bx_0$, on peut donc obtenir un intervalle probable pour Y_0. Cet intervalle sera d'autant plus grand que x_0 sera éloigné de $\bar{x}$.

Ainsi, pour notre exemple, on trouve dans la table que $P(|T_{n-2}| < 2{,}074) = 0{,}95$.

En prenant $x_0 = 100$, on a $y_0^* = 428{,}53$.

$$\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = 43{,}84\sqrt{1 + 0{,}06 + 0{,}03} = 45{,}15$$

d'où :

$$\left|\frac{Y_0 - 428{,}53}{45{,}15}\right| < 2{,}074$$

l'intervalle de prévision à 95 % est donc $334{,}89 < Y_0 < 522{,}17$, ce qui est assez imprécis malgré un coefficient de corrélation très élevé.

La variance de l'erreur de prévision dépend de deux termes : la variabilité intrinsèque de la variable Y_0 qui est égale à σ^2 et la variabilité dûe à l'imprécision des estimations de α et β dans la formule de régression qui dépend pour l'essentiel de la taille de l'échantillon et peut donc être réduite contrairement à la première source de variabilité.

La figure 16.5 montre la droite des moindres carrés encadrée par les deux types de contour à 95 % (ce sont des arcs d'hyperboles).

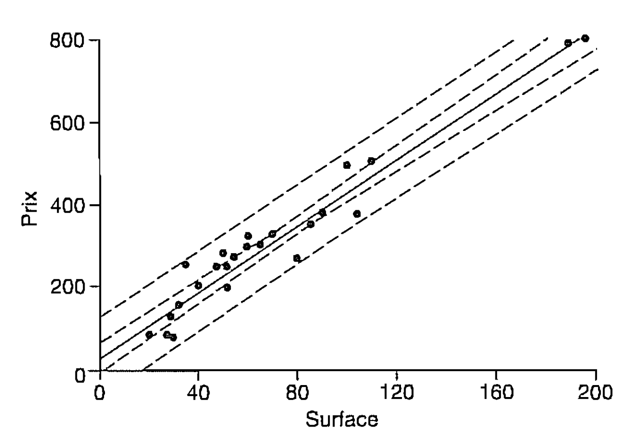


FIGURE 16.5 Régression du prix sur la surface.

Les limites les plus étroites correspondent à l'intervalle de confiance de la valeur moyenne $E(Y/X = x)$:

$$y^* \pm t\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

Les limites les plus éloignées correspondent à l'intervalle de prédiction pour une valeur unique :

$$y^* \pm t\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

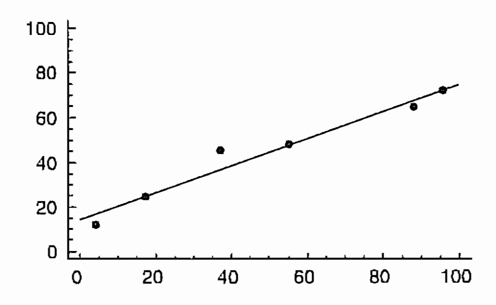
t est la valeur de la variable de Student telle que $P(|T_{n-2}| > t) = 5\ \%$.

16.5 UNE MÉTHODE DE RÉGRESSION ROBUSTE

La méthode des moindres carrés est sensible à la présence de données aberrantes situées loin de la droite de régression.

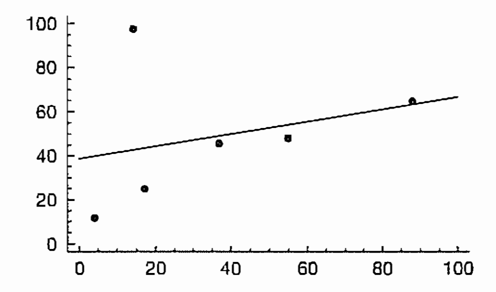
Ainsi sur l'exemple suivant, on a $y^* = 14{,}55 + 0{,}59X$ avec $r = 0{,}978$

i	y	x
1	11,797	4,1379
2	24,719	17,241
3	45,505	37,241
4	47,752	55,172
5	64,606	88,275
6	71,348	95,862



Si l'on modifie le dernier point en le remplaçant par $y = 97{,}191 \quad x = 14{,}482$

L'équation devient $y^* = 38{,}41 + 0{,}28\,x$ et le coefficient de corrélation tombe à $r = 0{,}29$.



On peut remédier à ce problème de deux façons :

– en éliminant les points « aberrants » ;

– en utilisant un autre critère que les moindres carrés.

La première solution peut être risquée et n'a de valeur que s'il s'agit effectivement de données erronées ou appartenant à une autre population, ce qui n'est pas toujours simple à déterminer.

La deuxième approche a l'avantage d'être automatique et de fournir un modèle robuste convenant à la majorité des données, en perdant toutefois les propriétés d'optimalité des estimateurs des moindres carrés, mais ces propriétés ne sont valables que sous certaines conditions.

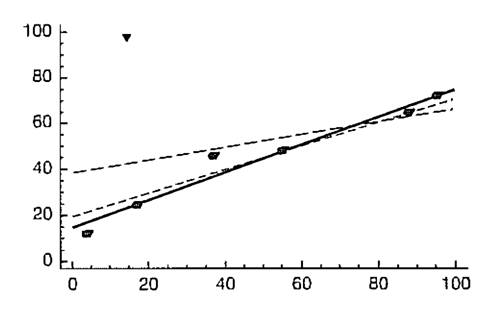
De nombreuses techniques robustes ont été proposées ; l'une des plus anciennes (elle a été proposé en 1757, soit 50 ans avant la méthode des moindres carrés) est la régression dite L_1 où on minimise la somme des valeurs absolues :

$$\sum_{i=1}^{n} |y_i^* - a - bx_i|$$

mais son utilisation a longtemps été négligée car contrairement aux moindres carrés, il n'existe pas de formule donnant les coefficients et leurs erreurs standard. Des algorithmes spécifiques sont nécessaires.

Dans l'exemple précédent cette méthode fournit l'équation $y^* = 19{,}66 + 0{,}51x$ et le graphique suivant montre que la solution L_1 (en pointillé), avec la donnée perturbée reste plus proche de la solution initiale (en gras) que celle des moindres carrés (en tireté large).

Une particularité de la régression L_1 est que la droite optimale passe toujours par deux des points de l'échantillon (ici les points 4 et 5) mais on ne peut savoir à l'avance lesquels. On pourra consulter Birkes et Dodge (1993) pour de plus amples développements.



16.6 RÉGRESSION NON PARAMÉTRIQUE

Lorsque la forme de la courbe de régression est complètement inconnue, on peut utiliser une estimation non paramétrique de la courbe $f(x) = E(Y/X = x)$ d'une manière semblable à l'estimation non-paramétrique de la densité (*cf.* 13.9).

En se donnant un intervalle $[x - h/2 \, ; x + h/2]$ centré sur x et de longueur h, le régressogramme consiste à compter le nombre de points appartenant à l'intervalle et à calculer la moyenne des y correspondants.

De manière analogue à l'estimateur de la fenêtre mobile, on peut écrire cette estimation :

$$\hat{E}(Y/X = x) = \frac{\sum_{i=1}^{n} K\left[\frac{x - x_i}{h}\right] y_i}{\sum_{i=1}^{n} K\left[\frac{x - x_i}{h}\right]} \quad \text{avec} \quad K(u) = 1 \quad \text{si} \quad -\frac{1}{2} \leq u \leq \frac{1}{2}$$

On obtient l'estimateur de Nadaraya-Watson en utilisant un noyau K continu, d'où une estimation continue de la fonction de régression. Son caractère plus ou moins lisse dépend de h que l'on peut optimiser au moyen d'une méthode de validation croisée : on cherche la valeur h qui minimise la somme des carrés des écarts en omettant à chaque fois dans la formule de Nadaraya-Watson la valeur x_i quand on fait l'estimation en ce point.

L'estimation non paramétrique de la régression fournit seulement des valeurs point par point. Cela peut paraître gênant si l'on cherche un modèle explicite, mais si l'on ne cherche pas à extrapoler en dehors du domaine observé de la variable explicative, on a toujours une prévision.

Sur des données de B. W. Silverman reliant l'accélération Y subie par le crâne d'un motocycliste en fonction du temps X après l'impact, on voit clairement l'intérêt de la méthode, car il n'y a pas de modèle simple pour la courbe de régression (calculs effectués avec XploRe 4.2).

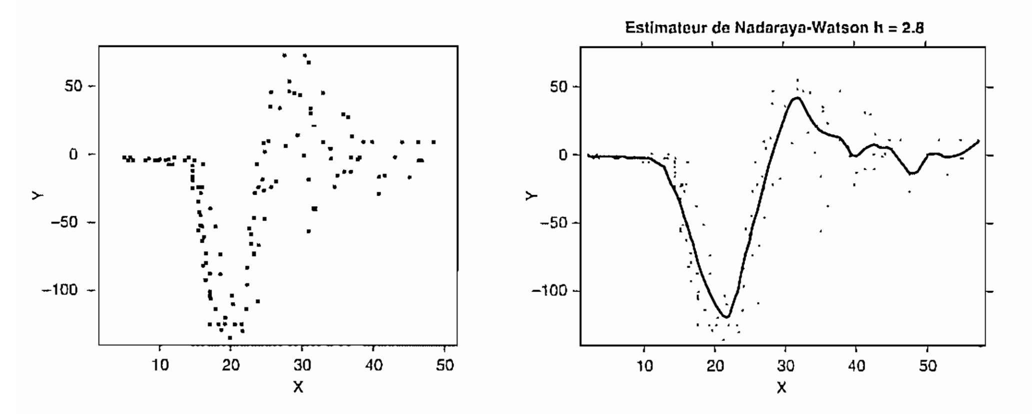


FIGURE 16.6

17 La régression multiple et le modèle linéaire général

Ce chapitre généralise le précédent. L'accent y est mis sur les interprétations géométriques. La complexité du sujet ne provient cependant pas tant de la difficulté des calculs, mais plutôt de la diversité des approches possibles que nous résumerons ici par la distinction entre modèle linéaire et régression multiple.

L'apparente simplicité d'utilisation des programmes de calcul, qui servent aussi bien pour la régression que pour le modèle linéaire, car les formules de résolution sont en pratique les mêmes, masque en réalité de profondes différences quant au modèle utilisé.

La pratique de la régression multiple est assez délicate comme l'illustreront les paragraphes 17.3 et 17.4.

17.1 RÉGRESSION ET MODÈLE LINÉAIRE

17.1.1 Régression entre variables aléatoires

17.1.1.1 Aspect empirique : la recherche d'un ajustement linéaire

On a mesuré sur n individus $p+1$ variables représentées par des vecteurs de $\mathbb{R}^n$ $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$; $\mathbf{y}$ est la variable à expliquer ou critère, les $\mathbf{x}_j$ les variables explicatives ou prédicteurs.

Les variables explicatives seront généralement supposées être linéairement indépendantes, ce qui ne veut pas dire qu'elles sont statistiquement indépendantes (en particulier, elles peuvent être corrélées). Il faut donc ***proscrire*** absolument la terminologie utilisée dans certains ouvrages où $\mathbf{y}$ est dite variable dépendante et les $\mathbf{x}_j$ variables indépendantes.

On cherche alors à reconstruire $\mathbf{y}$ au moyen des $\mathbf{x}_j$ par une formule linéaire.

On pose $\mathbf{y}^* = b_0\mathbf{1} + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_p\mathbf{x}_p$ et on désire que $\mathbf{y}^*$ soit le plus proche possible de $\mathbf{y}$.

Si l'espace des variables $\mathbb{R}^n$est muni comme d'habitude de la métrique $\mathbf{D}$, on exigera que $\|\mathbf{y} - \mathbf{y}^*\|^2$ soit minimal : c'est le critère des moindres carrés.

$\mathbf{y}^*$ est alors la projection $\mathbf{D}$-orthogonale de $\mathbf{y}$ sur le sous-espace W (de dimension $(p+1)$ en général) engendré par les variables $\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ (fig. 17.1).

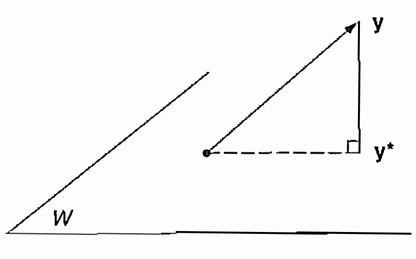


FIGURE 17.1

Soit **X** la matrice à n lignes dont les colonnes sont $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p$:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix} \quad \text{(on exigera } n > p\text{)}$$

On sait que l'opérateur de projection **D**-orthogonal sur W a pour expression : $\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$.

Donc :

$$\boxed{\mathbf{y}^* = \mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{y}}$$

En posant $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_p \end{bmatrix}$ on a : $\mathbf{y}^* = \mathbf{X}\mathbf{b}$ par hypothèse, donc :

$$\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{y}}$$

En particulier, si $\mathbf{D} = \frac{1}{n}\mathbf{I}$:

$$\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}} \quad \text{et} \quad \boxed{\mathbf{y}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

b est appelé vecteur des coefficients de régression.

17.1.1.2 Modèle probabiliste : l'hypothèse de régression linéaire multiple

Si l'on veut justifier autrement que par sa simplicité l'ajustement linéaire de **y** par les $\mathbf{x}_j$, on peut utiliser le modèle probabiliste suivant :

On suppose que $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ constituent un n échantillon d'observations indépendantes de $p + 1$ variables aléatoires $\psi, \varphi_1, \varphi_2, \ldots, \varphi_p$.

On sait que la recherche de la meilleure approximation de ψ par une fonction des φ_j est donnée par l'espérance conditionnelle $E[\psi/\varphi_1, \varphi_2, \ldots, \varphi_p]$.

On pose alors l'***hypothèse de régression linéaire multiple*** :

$$\boxed{E[\psi/\varphi_1, \varphi_2, \ldots, \varphi_p] = \beta_0 + \sum_{j=1}^{p} \beta_j \varphi_j}$$

qui conduit au modèle $\psi = \beta_0 + \sum_{j=1}^{p} \beta_j \varphi_j + \varepsilon$ où ε est une variable aléatoire d'espérance nulle non corrélée avec les φ_j. On note σ^2 la variance de ε.

En règle générale, les coefficients $\beta_0, \beta_1, \ldots, \beta_p$ et σ^2 sont inconnus ; il s'agit donc de les estimer le mieux possible.

Entre les réalisations $y_i, x_{i1}, \ldots, x_{ip}, e_i$ de $\psi, \varphi_1, \ldots, \varphi_p, \varepsilon$ il existe la relation suivante, déduite de l'hypothèse de régression linéaire multiple :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \qquad i = 1, 2, \ldots, n$$

ce qui s'écrit matriciellement $\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}}$ avec :

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

Sous des hypothèses assez générales, on montrera au paragraphe 17.2 que le vecteur **b** obtenu par la méthode des moindres carrés est la meilleure estimation du vecteur $\boldsymbol{\beta}$ et que l'on peut déduire simplement de $\|\mathbf{y} - \mathbf{y}^*\|^2$ la meilleure estimation sans biais de σ^2 qui sera :

$$\boxed{\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{y}^*\|^2}{n - p - 1}}$$

17.1.2 Le modèle linéaire général

17.1.2.1 Aspect empirique

Supposons que pour chaque ligne de X on ait k repétitions indépendantes de y.

On a donc un nuage de k vecteurs $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$ dans $\mathbb{R}^n$ (fig. 17.2).

On obtient ce genre de situations lorsque l'on fixe par exemple certaines valeurs des conditions expérimentales (température, pression, etc.) et qu'on mesure plusieurs

fois de suite un phénomène pour les mêmes combinaisons de valeurs des conditions expérimentales.

Le modèle linéaire consiste alors à postuler que le centre de gravité du nuage des $\mathbf{y}_1$; $\mathbf{y}_2$, . . . ; $\mathbf{y}_k$ se trouve dans W : $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$.

Le problème est alors le suivant : comment, à l'aide d'une seule observation $\mathbf{y}$, approximer le mieux possible $\mathbf{g}$? En effet, en réalité on ne connaît la plupart du temps qu'un seul point du nuage.

L'approximation $\mathbf{g}^*$ de $\mathbf{g}$ obtenue grâce à $\mathbf{y}$ peut s'exprimer comme la projection orthogonale de $\mathbf{y}$ sur W, selon une certaine métrique $\mathbf{M}$. Il faut alors choisir cette métrique $\mathbf{M}$ de telle sorte que $\mathbf{g}^*$ soit le plus proche possible de $\mathbf{g}$. Autrement dit, si l'on répétait l'opération de projection avec $\mathbf{y}_1$, $\mathbf{y}_2$, . . ., $\mathbf{y}_k$, les k approximations $\mathbf{g}_1^*$, $\mathbf{g}_2^*$, ..., $\mathbf{g}_k^*$, devraient être le plus concentrées possible autour de $\mathbf{g}$ avec $\mathbf{g}_i^* = \mathbf{X}(\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbf{y}_i$ (fig. 17.3).

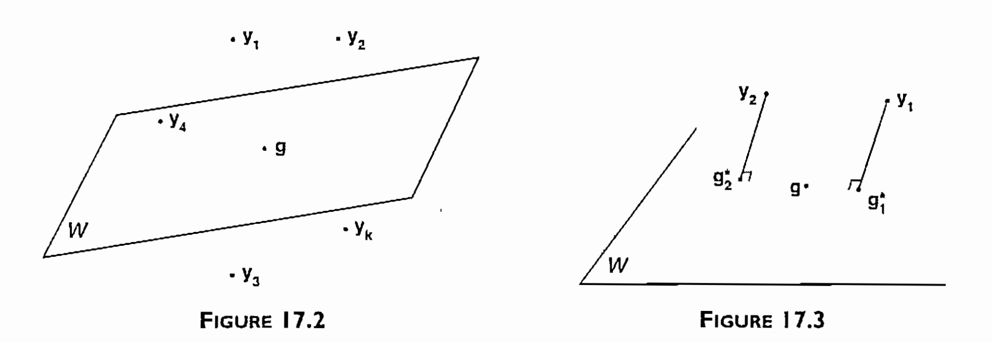


FIGURE 17.2

FIGURE 17.3

Il faut trouver la métrique M telle que l'inertie du nuage des $\mathbf{g}_i^*$ soit minimale.

Si $\mathbf{V}$ est la matrice de variance-covariance du nuage des $\mathbf{y}_i$, on démontre alors que ***la métrique M rendant l'inertie des $\mathbf{g}_i^*$ minimale est la métrique V^{-1}.***

Ce résultat constitue le théorème de Gauss-Markov généralisé[1].

Comme $\mathbf{g}_i^*$ est de la forme $\mathbf{X}\mathbf{b}_i$, ceci entraîne alors que le nuage des $\mathbf{b}_i$ est le moins dispersé possible dans $\mathbb{R}^{p+1}$, car la matrice de variance des $\mathbf{b}_i$ est égale à $(\mathbf{X}'\mathbf{X})$ fois celle des $\mathbf{g}_i^*$.

Avec une seule observation $\mathbf{y}$, on déduit :

$$\boxed{\begin{aligned} \mathbf{g}^* &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{b} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \end{aligned}}$$

1 Pour une démonstration complète on consultera l'ouvrage de Cailliez et Pagès, *Introduction à l'analyse des données*, p. 323 à 327.

17.1.2.2 Modèle probabiliste

Ce modèle n'est que la généralisation du cas précédent pour une infinité de répétitions.

On suppose que $\mathbf{y}$ est une réalisation d'un vecteur aléatoire d'espérance $\mathbf{X\beta}$ et de matrice variance Σ. Ceci revient à poser le modèle $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$ où $\mathbf{e}$ est une réalisation d'un vecteur aléatoire centré de matrice de variance Σ.

Le problème est alors d'estimer au mieux $\boldsymbol{\beta}$.

Suivant la notation de C. R. Rao, nous noterons en abrégé un tel modèle par le triplet $(\mathbf{y} ; \mathbf{X\beta} ; \Sigma)$.

On montre alors que le vecteur $\mathbf{b} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$ est, sous des hypothèses assez larges, l'estimation de variance minimale de $\boldsymbol{\beta} \cdot \mathbf{b}$ est appelé estimation des moindres carrés généralisés, par opposition à $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ estimation des moindres carrés ordinaires (mco).

17.1.3 Synthèse

Dans les deux cas, régression linéaire et modèle linéaire, on a été amené à poser le même modèle : $\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$.

Cependant, les hypothèses sont différentes : dans le modèle linéaire $\mathbf{X}$ est un tableau de données certaines, alors qu'en régression $\mathbf{X}$ est aléatoire.

Le vecteur des résidus $\mathbf{e}$ a une matrice variance quelconque Σ dans le modèle linéaire, alors qu'en régression le vecteur $\mathbf{e}$ a pour matrice variance $\sigma^2\mathbf{I}$ car l'hypothèse d'échantillonnage suppose les observations indépendantes.

Les objectifs sont également différents ; en régression, on veut ajuster au mieux $\mathbf{y}$; dans le modèle linéaire, on cherche à estimer l'effet moyen des variables explicatives.

Si l'on considère dans le modèle de régression linéaire multiple les variables explicatives comme des constantes, ce qui revient à travailler conditionnellement aux φ_j, il est clair que ceci revient au même que de poser le modèle linéaire $(\mathbf{y} ; \mathbf{X\beta} ; \sigma^2\mathbf{I}_n)$ si tous les individus ont le même poids.

En fait, la plupart des propriétés de la régression multiple s'obtiennent conditionnellement aux variables explicatives comme en régression simple, ce qui nous autorisera à ne plus parler que du modèle $(\mathbf{y} ; \mathbf{X\beta} ; \sigma^2\mathbf{I})$.

Par ailleurs, l'utilisation complète du modèle linéaire suppose connue la matrice Σ. Or, en pratique, on ignore Σ et, faute de mieux, on fait couramment l'hypothèse simplificatrice que Σ est diagonale (non corrélation des erreurs) et que tous les termes sont égaux (homoscédasticité), c'est-à-dire que $\Sigma = \sigma^2\mathbf{I}_n$, quitte à vérifier *a posteriori* sur les résultats la validité de ces deux hypothèses.

Ceci explique la confusion entre modèle linéaire et régression multiple ; dans ce qui suit, nous ne ferons plus la distinction, car nous nous référerons désormais à l'unique modèle simplificateur $(\mathbf{y} ; \mathbf{X\beta} ; \sigma^2\mathbf{I})$, en supposant que les poids des observations $p_i = 1/n$ sont égaux entre eux.

Remarquons pour finir que le terme de linéaire s'applique en fait au vecteur $\boldsymbol{\beta}$ et non aux variables explicatives ; ainsi, la régression polynomiale $\psi = \beta_0 + \beta_1\varphi + \beta_2\varphi^2 + \cdots \beta_p\varphi^p$ est un cas particulier du modèle général où l'on prend p variables explicatives $\varphi, \varphi^2, \ldots, \varphi^p$.

17.2 ESTIMATIONS ET TESTS DES PARAMÈTRES DU MODÈLE $(\mathbf{y} ; \mathbf{X}\boldsymbol{\beta} ; \sigma^2\mathbf{I})$

17.2.1 Estimation de $\boldsymbol{\beta}$ et de σ^2

17.2.1.1 Propriétés générales

Soit **b** la solution des moindres carrés : $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

THÉORÈME 1

b *est un estimateur sans biais de* $\boldsymbol{\beta}$.

Démonstration : $E(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y})$ car **X** est constante et est un opérateur linéaire. $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ par hypothèse du modèle linéaire général (ε, donc **e** est d'espérance nulle). Donc :

$$E(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

THÉORÈME 1 (DE GAUSS-MARKOV)

b *est de tous les estimateurs sans biais de* $\boldsymbol{\beta}$ *de la forme* **By**, *celui de variance minimale dans le sens qui sera précisé plus loin.*

Démonstration :

– La matrice variance de **b** est en effet $\boxed{V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}}$ car :

$$V(\mathbf{b}) = V((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{et} \quad V(\mathbf{y}) = V(\mathbf{e}) = \sigma^2\mathbf{I}_n.$$

– Soit **By** un autre estimateur linéaire de $\boldsymbol{\beta}$ sans biais.

Soit $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{B}\mathbf{y}$ la différence de ces deux estimateurs. Comme ils sont sans biais on a $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{B}\mathbf{X}\boldsymbol{\beta}$.

On a donc $\mathbf{B}\mathbf{X} = \mathbf{I}_{p+1}$ car cette relation doit être vérifiée pour tout $\boldsymbol{\beta}$.

Posons $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}$.

Comme $\mathbf{B}\mathbf{X} = \mathbf{I}_{p+1}$, on en déduit : $\mathbf{C}\mathbf{X} = \mathbf{0}$.

Cherchons la matrice de variances-covariances de **By** :

$$\begin{aligned} V(\mathbf{B}\mathbf{y}) &= \mathbf{B}V(\mathbf{y})\mathbf{B}' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}]\,\sigma^2\mathbf{I}_n[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C}]' \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}' + \mathbf{C}\mathbf{C}'] \end{aligned}$$

soit, puisque : $\mathbf{C}\mathbf{X} = \mathbf{0}$

$$= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{C}\mathbf{C}']$$

$$V(\mathbf{B}\mathbf{y}) = V(\mathbf{b}) + \sigma^2\mathbf{C}\mathbf{C}'$$

On en déduit que pour chaque composante de **b**, b_i est un estimateur meilleur que $(\mathbf{B}\mathbf{y})_i$ et que d'autre part $V(\mathbf{B}\mathbf{y}) - V(\mathbf{b})$ est semi-définie positive. (En effet, les termes diagonaux de $\mathbf{C}\mathbf{C}'$ sont ≥ 0).

Ce théorème est un cas particulier du théorème général énoncé en 17.1.2.1.

THÉORÈME

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{y}^*\|^2}{n-p-1} = \frac{\|\mathbf{y} - \mathbf{Xb}\|^2}{n-p-1}$$

est un estimateur sans biais de σ^2.

Démonstration : Considérons (voir fig. 17.4) le triangle rectangle dont les sommets sont les extrémités des vecteurs $\mathbf{y}$, $\mathbf{Xb}$ et $\mathbf{X\beta}$.

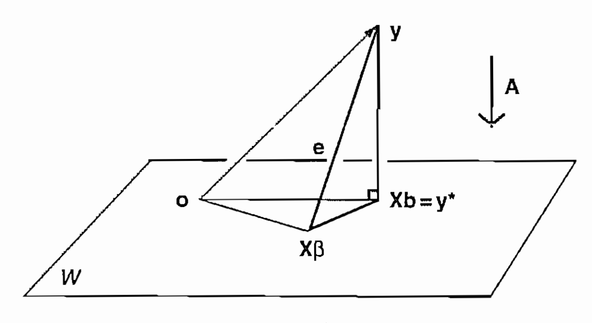


FIGURE 17.4

Soit $\mathbf{A}$ le projecteur sur $W(\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ et $\mathbf{I} - \mathbf{A}$ le projecteur sur $W^{\perp}$.

Puisque $\mathbf{y} - \mathbf{Xb}$ est orthogonal à W, on voit aisément que $\mathbf{y} - \mathbf{Xb}$ est égal à $(\mathbf{I} - \mathbf{A})\mathbf{e}$ car $\mathbf{e}$ se décompose sur W et $W^{\perp}$ en $\mathbf{Xb} - \mathbf{X\beta}$ et $\mathbf{y} - \mathbf{Xb}$ respectivement.

Donc $\mathbf{y} - \mathbf{Xb} = \mathbf{e}'(\mathbf{I} - \mathbf{A})'(\mathbf{I} - \mathbf{A})\mathbf{e}$ et comme $\mathbf{I} - \mathbf{A}$ est un projecteur :

$$(\mathbf{I} - \mathbf{A})' = \mathbf{I} - \mathbf{A} = (\mathbf{I} - \mathbf{A})^2$$

et on obtient alors :

$$\|\mathbf{y} - \mathbf{Xb}\|^2 = \mathbf{e}'(\mathbf{I} - \mathbf{A})\mathbf{e} = \sum_{i,j} \alpha_{ij} e_i e_j$$

où α_{ij} est le terme courant de $(\mathbf{I} - \mathbf{A})$.

Donc $E[\|\mathbf{y} - \mathbf{Xb}\|^2] = \sum_{i,j} \alpha_{ij} E(e_i e_j)$.

Comme les e_i sont non corrélés $E(e_i e_j) = \delta_{ij}\sigma^2$ où δ_{ij} est le symbole de Kronecker.

Donc $E[\|\mathbf{y} - \mathbf{Xb}\|^2] = \sigma^2 \sum_{i=1}^{n} \alpha_{ii} = \sigma^2 \operatorname{trace}(\mathbf{I} - \mathbf{A})$.

On sait que la trace d'un projecteur est égale à son rang (car ses valeurs propres sont 0 ou 1), c'est-à-dire à la dimension de l'espace d'arrivée qui est ici $W^{\perp}$. Comme $\dim W = p + 1$, on a $\dim W^{\perp} = n - p - 1$:

$$E[\|\mathbf{y} - \mathbf{Xb}\|^2] = \sigma^2(n - p - 1)$$

17.2.1.2 *Propriétés supplémentaires si e est gaussien*

Introduisons alors l'hypothèse $e_i \in LG(0\,;\sigma)\ \forall i$.

La densité du vecteur aléatoire **y** s'écrit :

$$L(\mathbf{y}, \boldsymbol{\beta}, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

car **y** est alors un vecteur gaussien multidimensionnel ; $\mathbf{y} \in N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

THÉORÈME

Les estimateurs de maximum de vraisemblance de $\boldsymbol{\beta}$ *et* σ^2 *sont :*

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{et} \quad \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \ (\textit{qui est biaisé})$$

La démonstration est laissée au soin du lecteur.

Conformément à ce qui a été développé dans la partie consacrée à l'estimation nous allons rechercher des statistiques exhaustives pour les paramètres inconnus $\boldsymbol{\beta}$ et σ^2 afin d'étudier l'optimalité des estimateurs associés, car la propriété du maximum de vraisemblance ne nous renseigne pas sur l'efficacité des estimateurs.

La densité de **y** peut s'écrire :

$$L(\mathbf{y}, \boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} - \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})\right)$$

Soit $\mathbf{a}(\boldsymbol{\beta}, \sigma^2)$ le vecteur ligne à $p + 2$ composantes :

$$\left(-\frac{1}{2\sigma^2}, \frac{\beta_0}{\sigma^2}, \frac{\beta_1}{\sigma^2}, \frac{\beta_2}{\sigma^2}, \ldots, \frac{\beta_p}{\sigma^2}\right)$$

et :

$$\mathbf{T}(\mathbf{y}) = \begin{bmatrix} \mathbf{y}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

On a :

$$L(\mathbf{y}, \boldsymbol{\beta}, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(\mathbf{a}(\boldsymbol{\beta}, \sigma^2)\mathbf{T}(\mathbf{y}) + \mathbf{C}(\boldsymbol{\beta}, \sigma^2)\right)$$

où :

$$\mathbf{C}(\boldsymbol{\beta}, \sigma^2) = -\frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}$$

Le domaine de définition de **y** ne dépendant pas de $\boldsymbol{\beta}$ ni de σ^2, et le rang de **X** étant $p + 1$, l'application définie par $\mathbf{T} = \begin{bmatrix} \mathbf{y}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$ est bijective. D'après le théorème de Darmois généralisé, $\mathbf{T}(\mathbf{y})$ constitue une statistique exhaustive.

b et $\hat{\sigma}^2$ qui sont fonction de **T** sont donc les estimateurs sans biais de variance minimale de $\boldsymbol{\beta}$ et σ^2.

De plus **b** transformé linéaire d'un vecteur gaussien est lui-même gaussien.

$$\boxed{\mathbf{b} \in N_{p+1}(\boldsymbol{\beta} \,;\, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)}$$

17.2.1.3 *Lois des côtés du triangle rectangle y, y*, Xβ (fig. 17.5)*

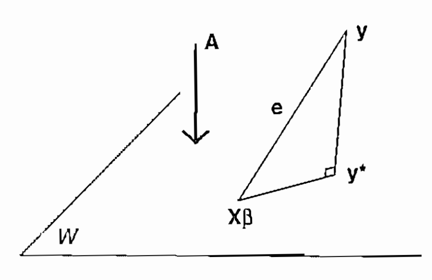


FIGURE 17.5

Ce triangle est rectangle en $\mathbf{y}^*$; le théorème de Pythagore s'écrit donc :

$$\|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{Xb}\|^2 + \|\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Or $\mathbf{y} - \mathbf{Xb} = (\mathbf{I} - \mathbf{A})\mathbf{e}$ et $\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta} = \mathbf{Ae}$. On a donc $\|\mathbf{e}\|^2 = \mathbf{e}'\mathbf{Ae} + \mathbf{e}'(\mathbf{I} - \mathbf{A})\mathbf{e}$. **e** est un vecteur gaussien où les e_i suivent indépendamment des lois LG(0 ; σ). Donc :

$$\frac{\|\mathbf{e}\|^2}{\sigma^2} = \sum_i e_i^2/\sigma^2 = \chi^2_n$$

Le théorème de Pythagore se transforme alors en théorème de Cochran et on trouve que:

$$\boxed{\frac{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{Xb}\|^2}{\sigma^2} \text{ suit un } \chi^2_{p+1}} \qquad \boxed{\frac{\|\mathbf{y} - \mathbf{Xb}\|^2}{\sigma^2} \text{ suit un } \chi^2_{n-p-1}}$$

et ces deux variables sont indépendantes comme formes quadratiques de rang $p + 1$ et $n - p - 1$ (rangs de projecteurs) de n variables normales centrées-réduites.

On peut ainsi obtenir des intervalles de confiance pour σ.

17.2.1.4 *Le modèle (y ; Xβ ; Σ)*

Par les mêmes procédés on peut montrer que :

1) $\mathbf{b} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$ est l'estimateur de variance minimale de **β** parmi les estimateurs fonctions linéaires de **y** (théorème de Gauss-Markov généralisé).
2) Si l'hypothèse de normalité est vérifiée, **b** est l'estimateur du maximum de vraisemblance et est de variance minimale.

17.2.2 Tests dans le modèle linéaire

17.2.2.1 Le coefficient de corrélation multiple R et l'analyse de variance de la régression

R est le coefficient de corrélation entre la série $y_1, y_2, \ldots, y_n$ et la série $y_1^*, y_2^*, \ldots, y_n^*$. En d'autres termes, c'est la valeur maximale du coefficient de corrélation linéaire simple entre les coordonnées de $\mathbf{y}$ et les coordonnées de tout vecteur de la forme $\mathbf{Xb}$ (voir chapitre 6).

Comme tout coefficient de corrélation linéaire, son carré s'interprète en termes de variance expliquée :

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2} = \frac{\text{Variance expliquée par la régression}}{\text{Variance des } y}$$

Si $R^2 = 1$, $y_i = y_i^* \ \forall i$ l'ajustement est parfait.

R^2 est appelé coefficient de détermination.

Géométriquement R est le cosinus de l'angle formé par $\mathbf{y} - \bar{\mathbf{y}}$ et $\mathbf{y}^* - \bar{\mathbf{y}}$ dans $\mathbb{R}^n$ où $\bar{\mathbf{y}}$ est le vecteur dont toutes les composantes sont égales à $\bar{y}$. $\bar{\mathbf{y}}$ est la projection de $\mathbf{y}$ sur la droite des constantes qui appartient à W (fig. 17.6). Voir chapitre 6 (§ 6.2).

$\bar{\mathbf{y}}$

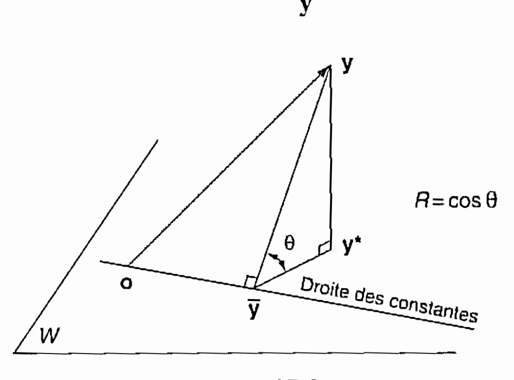


FIGURE 17.6

D'après le théorème des trois perpendiculaires,   est aussi la projection de $\mathbf{y}^*$ sur la droite des constantes.

Le coefficient R^2 est utilisé pour tester la qualité de l'ajustement de $\mathbf{y}$ par $\mathbf{y}^*$ (analyse de variance de la régression).

Il est facile d'écrire la décomposition classique :

$$\underbrace{\frac{1}{n}\sum_i (y_i - \bar{y})^2}_{\text{Variance totale}} = \underbrace{\frac{1}{n}\sum_i (y_i - y_i^*)^2}_{\text{Variance résiduelle}} + \underbrace{\frac{1}{n}\sum_i (y_i^* - \bar{y})^2}_{\text{Variance expliquée par la régression}}$$

On sait que $\frac{\sum_i (y_i - y_i^*)^2}{\sigma^2}$ suit $\forall \boldsymbol{\beta}$ une loi χ^2_{n-p-1} et on peut montrer que $\frac{\sum_i (y_i^* - \bar{y})^2}{\sigma^2}$ est un χ^2_p si $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ (β_0 quelconque).

Si $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, alors $\frac{1}{\sigma^2}\sum_i (y_i - \bar{y})^2$ suit un χ^2_{n-1} comme variance d'un échantillon de variables normales de mêmes lois.

Comme $\frac{R^2}{1 - R^2} = \frac{\sum_i (y_i^* - \bar{y})^2}{\sum_i (y_i - y_i^*)^2}$, on trouve que si $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ (mais β_0 quelconque) :

$$\boxed{\frac{R^2}{1 - R^2}\,\frac{n - p - 1}{p} = F(p\,; n - p - 1)}$$

On retrouve comme cas particulier la loi du coefficient de corrélation usuel si $p = 1$.

Le test du R^2 est le même que celui de la nullité de q coefficients de régression lorsque $q = p$ (voir paragr. 17.2.2.3).

L'hypothèse de non-régression $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ correspond à la nullité de coefficient de corrélation multiple théorique $\mathcal{R}$ dans le cadre de la régression entre variables aléatoires.

Sous cette hypothèse nulle la loi de $\mathcal{R}$ est celle d'une variable bêta de type I de paramètre $\frac{p}{2}$ et $\frac{n - p - 1}{2}$ on en déduit que $E(R^2) = \frac{p}{n - 1}$ et $V(R^2) = \frac{2(n - p - 1)p}{(n^2 - 1)(n - 1)}$.

Si l'hypothèse de non-régression n'est pas satisfaite ($\mathcal{R}^2 \neq 0$), la loi de R^2 ne prend pas une forme aussi simple et R^2 est alors un estimateur biaisé de $\mathcal{R}^2$.

On montre en effet que $E(R^2) = \mathcal{R}^2 + \frac{p}{n - 1}(1 - \mathcal{R}^2) + 0\left(\frac{1}{n^2}\right)$ d'où la définition du $\boldsymbol{R^2}$ **ajusté** $\hat{R}^2$:

$$\boxed{\hat{R}^2 = \frac{(n - 1)\,R^2 - p}{n - p - 1}}$$

où le biais en $1/n$ est éliminé mais qui peut conduire à des valeurs négatives si $\mathcal{R}^2$ est voisin de 0.

Un calcul élémentaire montre que $\hat{\sigma}^2 = \frac{n}{n - 1}(1 - \hat{R}^2)s_y^2$.

17.2.2.2 Test du caractère significatif d'un des coefficients de régression

Il s'agit de tester $\beta_j = 0$ contre $\beta_j \neq 0$.

Soit b_j le coefficient de régression empirique. On sait que $V(b_j) = \sigma^2\,[(\mathbf{X}'\mathbf{X})^{-1}_{j,j}]$ où $[(\mathbf{X}'\mathbf{X})^{-1}_{j,j}]$ est le terme (j, j) de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$.

Comme $\sum_i \frac{(y_i - y_i^*)^2}{\sigma^2}$ suit un χ^2_{n-p-1}, il vient immédiatement que :

$$\frac{(b_j - \beta_j)}{\sqrt{\sum_i (y_i - y_i^*)^2[(\mathbf{X}'\mathbf{X})^{-1}_{j,j}]}}\sqrt{n-p-1}$$

suit un t de Student à $n - p - 1$ degrés de liberté, ce qui permet de tester l'hypothèse $\beta_j = 0$, car b_j suit une loi $LG(\beta_j, \sigma\sqrt{((\mathbf{X}'\mathbf{X})^{-1}_{jj}]})$.

On peut aussi écrire :

$$t_{n-p-1} = \frac{(b_j - \beta_j)}{\sqrt{\frac{\|\mathbf{y} - \mathbf{Xb}\|^2}{n-p-1}[(\mathbf{X}'\mathbf{X})^{-1}_{jj}]}}$$

Le dénominateur de l'expression précédente est appelé erreur standard ou écart-type du coefficient de régression et désigné par la lettre s_j. (Il s'agit en fait de la racine carrée de l'estimation sans biais de la variance de b_j).

On prendra garde au fait que les statistiques de test des coefficients de régression ne sont pas indépendantes car les b_j ne sont pas indépendants. On peut par exemple trouver un R^2 significatif sans qu'aucun coefficient de régression pris isolément soit significativement différent de zéro (c'est souvent le cas lorsque les prédicteurs sont fortement corrélés entre eux, voir exemple plus loin).

17.2.2.3 Test de q coefficients de régression, test d'une sous-hypothèse linéaire

Les deux tests précédents ne sont en fait que des cas particuliers du test plus général suivant qui permet, entre autres choses, de tester la nullité de q coefficients de régression.

Écrire $\beta_1 = \beta_{10}, \beta_2 = \beta_{20}, \ldots, \beta_q = \beta_{q0}$ n'est qu'un cas particulier de $\mathbf{H}\boldsymbol{\beta} = \boldsymbol{\theta}$, où $\mathbf{H}$ est une matrice de rang q.

Le test de $H_0 : \mathbf{H}\boldsymbol{\beta} = \boldsymbol{\theta}$ contre $H_1 : \mathbf{H}\boldsymbol{\beta} \neq \boldsymbol{\theta}$ s'effectue alors de la manière suivante : on pose $\mathbf{y}^*$ la solution des moindres carrés $\mathbf{y}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ et $\mathbf{y}_0^*$ la solution des moindres carrés sous la contrainte $\mathbf{H}\boldsymbol{\beta} = \boldsymbol{\theta}$ (on projette sur le sous-espace de W vérifiant cette contrainte).

On montre alors que si H_0 est vraie :

$$\frac{\|\mathbf{y} - \mathbf{y}_0^*\|^2 - \|\mathbf{y} - \mathbf{y}^*\|^2}{\|\mathbf{y} - \mathbf{y}^*\|^2}\frac{n-p-1}{q} = F(q\,;n-p-1)$$

ce qui permet de tester H_0.

Ce test a pour cas particulier le ***test simultané de tous les coefficients de régression*** $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ contre $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

Comme $\|\mathbf{X\beta} - \mathbf{Xb}\|^2$ et $\|\mathbf{y} - \mathbf{Xb}\|^2$ sont indépendantes, on en déduit que :

$$\frac{\|\mathbf{X\beta} - \mathbf{Xb}\|^2}{\|\mathbf{y} - \mathbf{Xb}\|^2} \frac{n-p-1}{p+1} = F(p+1, n-p-1)$$

En remplaçant $\boldsymbol{\beta}$ par $\boldsymbol{\beta}_0$ on peut donc tester l'hypothèse nulle. On rejettera H_0 si la quantité $\frac{\|\mathbf{X\beta}_0 - \mathbf{y}^*\|^2}{\|\mathbf{y} - \mathbf{y}^*\|^2} \frac{n-p-1}{p+1}$ est trop grande.

Remarque : Ce dernier test suppose également une valeur *a priori* pour β_0. Ce n'est pas le test le plus couramment utilisé (qui suppose β_0 inconnu).

17.2.3 Intervalle de prévision pour une valeur future

Cherchons à encadrer la valeur prévue y_0^* pour un individu supplémentaire pour lequel les variables explicatives prennent les valeurs $x_{10}, x_{20}, \ldots, x_{p0}$.

Posons :
$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{10} \\ \cdot \\ \cdot \\ \cdot \\ x_{p0} \end{bmatrix}$$

alors $\mathbf{y}_0^* = \mathbf{x}_0'\mathbf{b}$ est une variable aléatoire suivant une loi $LG(\mathbf{x}_0'\boldsymbol{\beta}_0 ; \sigma\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0})$ d'après les résultats usuels sur les combinaisons linéaires de variables gaussiennes.

Comme au chapitre précédent, par studentisation, puisque σ doit être estimé, il vient :

$$\boxed{\frac{y_0 - y_0^*}{\hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} = T_{n-p-1}}$$

ce qui permet d'encadrer la « vraie valeur » y_0.

L'intervalle de confiance pour $E(Y/X = x_0)$ s'obtient en enlevant le 1 sous le radical.

17.3 L'ANALYSE DES RÉSULTATS

Les principaux problèmes abordés ici concernent la stabilité des résultats d'une régression. On distinguera les questions relatives à l'influence d'observations particulières et celles relatives à l'influence des variables sur les estimations (multicolinéarité). L'analyse des résidus est également un moyen de vérifier les hypothèses de base du modèle.

17.3.1 L'étude des résidus et des observations influentes

L'étude des résidus $y_i - y_i^*$ est fondamentale à plus d'un titre : elle permet tout d'abord de repérer des observations éventuellement aberrantes ou des observations qui jouent un rôle

important dans la détermination de la régression. Ensuite l'étude des résidus est bien souvent la seule façon de vérifier empiriquement le bien-fondé des hypothèses du modèle : linéarité, homoscédasticité, etc. : les graphes des résidus en fonction des variables explicatives ne doivent laisser apparaître aucune tendance.

Il est facile d'obtenir la matrice de variance des résidus puisque $\mathbf{y} = \mathbf{y} - \mathbf{Xb} + \mathbf{Xb}$ où $\mathbf{y} - \mathbf{Xb}$ est orthogonal à $\mathbf{Xb}$ d'où $V(\mathbf{y}) = V(\mathbf{y} - \mathbf{Xb}) + V(\mathbf{Xb})$ soit :

$$\sigma^2 \mathbf{I}_n = V(\mathbf{y} - \mathbf{Xb}) + \sigma^2 \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

d'où :

$$V(\mathbf{y} - \mathbf{Xb}) = \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$$

ce qui rappelle que les résidus sont en général corrélés entre eux.

En désignant par h_i le $i^{\text{ème}}$ terme diagonal du projecteur $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ on a :

$$V(y_i - y_i^*) = (1 - h_i)\sigma^2 \qquad \left(\text{où } \frac{1}{n} \leq h_i \leq 1 \qquad \text{avec} \sum_{i=1}^{n} h_i = p + 1\right)$$

d'où l'estimation de la variance du résidu :

$$\hat{V}(y_i - y_i^*) = \hat{\sigma}^2(1 - h_i)$$

On appelle résidu studentisé la quantité :

$$\frac{y_i - y_i^*}{\hat{\sigma}\sqrt{1 - h_i}}$$

Lorsque n est grand les résidus studentisés doivent rester compris entre -2 et 2.

Un fort résidu peut indiquer une valeur aberrante. Cependant une valeur peut être aberrante sans que son résidu soit important (voir fig. 17.7).

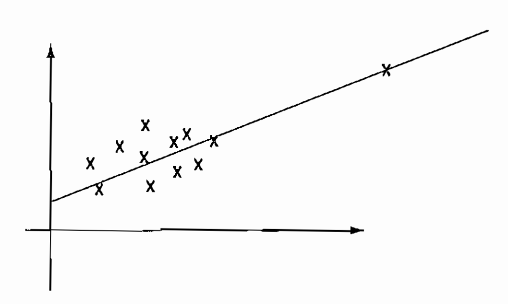


FIGURE 17.7

Il faut alors étudier l'influence de chaque observation sur les résultats.

On peut tout d'abord étudier l'influence d'une observation sur sa propre prédiction.

On appelle résidu prédit l'écart $y_i - y^*_{(-i)}$ où $y^*_{(-i)}$ est la prévision obtenue avec l'échantillon de $(n - 1)$ observations excluant la $i^{\text{ème}}$.

On peut vérifier que le résidu prédit vaut $\frac{y_i - y_i^*}{1 - h_i}$; il convient donc d'être prudent avec des observations dont le h_i serait grand. La quantité suivante notée *Press* est une mesure du pouvoir prédictif du modèle :

$$Press = \sum_{i=1}^{n} (y_i - y_{(-i)}^*)^2 = \sum_{i=1}^{n} \frac{(y_i - y_i^*)^2}{(1 - h_i)^2}$$

On peut enfin étudier l'influence d'une observation sur les estimations b_j des coefficients de régression et calculer par exemple une distance entre $\mathbf{b}$ et $\mathbf{b}_{(-i)}$ où $\mathbf{b}_{(-i)}$ est l'estimation de $\boldsymbol{\beta}$ obtenue sans la $i^{\text{ème}}$ observation.

La distance de Cook est l'une des plus utilisées :

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{(-i)})'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \mathbf{b}_{(-i)})}{(p + 1)\hat{\sigma}^2}$$

On montre que :

$$D_i = \frac{1}{p + 1} r_i^2 \frac{h_i}{1 - h_i} = \frac{\|\mathbf{y}^* - \mathbf{y}_{(-i)}^*\|^2}{(p + 1)\hat{\sigma}^2}$$

ou $\mathbf{y}_{(-i)}^* = \mathbf{X}\mathbf{b}_{(-i)}$.

Une distance D_i supérieure à 1 indique en général une influence anormale (*cf.* Cook-Weisberg, 1982).

17.3.2 La stabilité des coefficients de régression

L'écart-type s_j du coefficient b_j est déjà un indicateur du caractère plus ou moins stable de l'estimation d'un coefficient. Il est clair que si s_j est du même ordre de grandeur que b_j, ce dernier est mal déterminé.

La source principale d'instabilité dans l'estimation de $\boldsymbol{\beta}$ est la multicolinéarité : on désigne par cette expression la situation où les variables explicatives sont très corrélées entre elles.

Comme $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, si les prédicteurs sont très corrélés entre eux $\mathbf{X}'\mathbf{X}$ est mal conditionnée (déterminant proche de 0) et son inverse aura des termes élevés.

Dans ce cas les paramètres du modèle seront estimés avec imprécision et les prédictions pourront être entachés d'erreurs considérables même si R^2 a une valeur élevée.

Il est donc important de mesurer l'effet de la colinéarité entre les prédicteurs, cela s'effectue au moyen des facteurs d'inflation de la variance et des valeurs propres de la matrice de corrélation.

Il sera commode pour la suite de supposer que toutes les variables sont centrées et réduites (ce qui ne change pas le R^2 ni les valeurs prévues) et que l'on effectue donc une régression sans constante : $(\mathbf{X}'\mathbf{X})$ est donc une matrice de taille p et $\mathbf{b}$ un vecteur à p composantes. On a donc $(\mathbf{X}'\mathbf{X}) = n\mathbf{R}$ où $\mathbf{R}$ est la matrice de corrélation entre les prédicteurs.

17.3.2.1 Le facteur d'inflation de la variance (VIF)

On a donc : $$V(\mathbf{b}) = \sigma^2 \frac{\mathbf{R}^{-1}}{n} \quad \text{et} \quad V(b_j) = \frac{\sigma^2}{n}(\mathbf{R}^{-1})_{jj}.$$

Or $(\mathbf{R}^{-1})_{jj}$, j^e terme diagonal de $\mathbf{R}^{-1}$ n'est autre que $\dfrac{1}{1 - R_j^2}$ où R_j^2 est le carré du coefficient de corrélation multiple de $\mathbf{x}^j$ avec les $p - 1$ autres variables explicatives.

Si les p variables explicatives étaient orthogonales la régression multiple reviendrait à p régressions simples ; $V(b_j)$ serait égal à $\dfrac{\sigma^2}{n}$.

Le terme $\dfrac{1}{1 - R_j^2}$ est appelé « facteur d'inflation de la variance » tandis que $1 - R_j^2$ est appelé « tolérance ». La moyenne des p facteurs d'inflation est utilisée parfois comme indice global de multicolinéarité.

17.3.2.2 Le rôle des valeurs propres de R

Posons $\mathbf{R} = \mathbf{U\Lambda U}'$ où Λ est la matrice diagonale des valeurs propres et $\mathbf{U}$ la matrice des vecteurs propres de $\mathbf{R}$.

On a donc $\mathbf{R}^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}'$.

On en déduit : $$V(b_j) = \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{(u_{jk})^2}{\lambda_k}$$

On voit donc que $V(b_j)$ dépend des inverses de valeurs propres de $\mathbf{R}$: lorsqu'il y a forte colinéarité entre les prédicteurs les dernières valeurs propres sont proches de zéro d'où l'instabilité des b_j.

17.4 SÉLECTION DE VARIABLES

Plutôt que de chercher à expliquer $\mathbf{y}$ par toutes les p variables explicatives, on peut chercher seulement un ensemble de q variables parmi les p qui donne une reconstitution presque aussi satisfaisante de $\mathbf{y}$.

Les objectifs d'une telle démarche sont multiples : économiser le nombre de prédicteurs, obtenir des formules stables et d'un bon pouvoir prédictif en éliminant des variables redondantes qui augmentent le facteur d'inflation de la variance, obtenir un modèle plus facile à interpréter.

17.4.1 Les critères de choix

Ils dépendent bien sûr des usages que l'on fait de la régression : reconstitution des y_i, prévision de valeurs futures, ou estimation précise des paramètres d'un modèle.

Le critère du R^2 est bien adapté au premier objectif et est celui qui est le plus utilisé dans les programmes de régression pas à pas. Il n'est cependant pas à l'abri des critiques : il varie de façon monotone avec le nombre des variables : il ne peut qu'augmenter si l'on rajoute un

prédicteur même peu corrélé avec y puisque la dimension de W augmente. On ne peut donc l'utiliser pour choisir la taille d'un sous-ensemble de prédicteurs.

Si l'objectif est de minimiser l'erreur de prévision le R^2 n'est pas adapté et on préfèrera des critères tels que le $\hat{\sigma}^2$ ou le *Press*.

$\hat{\sigma}^2$ ne varie pas de façon monotone avec le nombre de variables car :

$$\hat{\sigma}^2 = \frac{n}{n-k-1}(1 - R^2)s_y^2$$

Par contre $\hat{\sigma}^2$ varie de façon monotone avec le R^2 ajusté $\hat{R}^2$. Il est donc plus intéressant de prendre $\hat{R}^2$ que R^2 comme critère de qualité, ce qui permet de comparer des formules de régression comprenant des nombres différents de variables et de choisir celle qui minimise $\hat{\sigma}^2$ (ou maximise $\hat{R}^2$). On peut également utiliser les critères, informationnels AIC et BIC, voir chapitre 19.

17.4.2 Les techniques de sélection

17.4.2.1 *Recherche exhaustive*

Lorsque p n'est pas trop grand on peut étudier toutes les formules possibles : il y a C_p^k formules à k variables et donc $2^p - 1$ régressions.

A p fixé on choisira celle qui fournit le R^2 maximum, et si p n'est pas fixé celle qui fournit le $\hat{\sigma}^2$ minimum, ou le minimum d'autres critères (voir chapitre 19, § 19.4).

17.4.2.2 *Les méthodes de pas à pas*

Elles sont utilisées lorsque p est élevé et qu'il n'est pas possible de procéder à une recherche exhaustive.

Elles procèdent par élimination successive ou ajout successif de variables.

La méthode descendante consiste à éliminer la variable la moins significative parmi les p : en général celle qui provoque la diminution la plus faible des R^2 (c'est celle qui a le t de Student le moins significatif). On recalcule alors la régression et on recommence jusqu'à élimination de $p - 1$ variables ou en fonction d'un test d'arrêt.

La méthode ascendante procède en sens inverse : on part de la meilleure régression à une variable et on ajoute celle qui fait progresser le plus le R^2.

La méthode dite *stepwise* est un perfectionnement de l'algorithme précédent qui consiste à effectuer en plus à chaque pas des tests de signification du type Student ou F pour ne pas introduire une variable non significative et pour éliminer éventuellement des variables déjà introduites qui ne seraient plus informatives compte tenu de la dernière variable sélectionnée. L'algorithme s'arrête quand on ne peut plus ajouter ni retrancher de variables.

Ces méthodes ne donnent pas forcément les meilleures régressions à k variables ni les mêmes résultats si l'on les emploie en concurrence, mais elles sont très pratiques d'emploi, la méthode *stepwise* semblant la meilleure. Elles ne mettent cependant pas à l'abri de l'élimination intempestive de variables réellement significatives, ce qui risque de biaiser les résultats. Il faut à ce propos rappeler que si l'on sait (par un modèle

physique par exemple) qu'une variable doit figurer dans un modèle, ce n'est pas parce qu'un test statistique la déclare non significative qu'il faut la rejeter (erreur de deuxième espèce).

17.5 TRAITEMENT DE LA MULTICOLINÉARITÉ

Lorsque les variables explicatives sont fortement corrélées entre elles, les variances des coefficients de régression deviennent très élevées : les estimations sont donc imprécises. En effet le déterminant de la matrice $\mathbf{X'X}$ est alors proche de 0, d'où des valeurs instables pour $V(\mathbf{b}) = \sigma^2(\mathbf{X'X})^{-1}$.

Ceci se produit entre autres si le nombre d'observations est faible devant le nombre de variables. Un cas extrême autrefois banni des manuels, est celui où le nombre de variables est supérieur au nombre d'observations : $\mathbf{X'X}$ n'est alors pas inversible.

Les méthodes de sélection de variables sont une des réponses possibles au problème de la multicolinéarité, mais elles peuvent conduire à l'élimination de variables significativement liées à y. Il est alors difficile de proposer à l'utilisateur un modèle qui ne tient pas compte de variables pourtant influentes et ne permet pas de quantifier l'effet de leurs variations sur la réponse y.

Les trois méthodes que nous présentons maintenant permettent de résoudre ce problème et même de traiter le cas où $p > n$. Ceci se fait au prix de la perte de certaines propriétés comme l'absence de biais des estimateurs et l'invariance par changement d'échelle : sur un plan technique on procédera à une standardisation préalable des variables par centrage-réduction.

D'après le théorème de Gauss-Markov, la méthode des moindres carrés fournit les estimateurs de variance minimale des β_j parmi les estimateurs sans biais. On ne pourra donc diminuer la variance des estimateurs qu'en utilisant des estimateurs biaisés. Comme l'erreur quadratique est égale à la variance plus le carré du biais, il est possible dans certaines conditions d'obtenir des estimations plus précises des coefficients avec un léger biais.

17.5.1 Régression sur composantes principales

D'après la formule établie au paragraphe 17.3.2.2 on diminuera $V(b_j)$ en ne retenant que certains termes de la somme des $\dfrac{(u_{jk})^2}{\lambda_k}$.

Ceci revient à la pratique suivante : on remplace les p variables explicatives par leurs p composantes principales qui engendrent le même espace W, et on effectue la régression sur les composantes principales ce qui revient à p régressions simples :

$$\mathbf{y}^* = \sum_{j=1}^{p} \alpha_j \mathbf{c}_j \qquad \text{où } \alpha_j = \frac{r(\mathbf{y}\,;\,\mathbf{c}_j)s_y}{\sqrt{\lambda_j}}$$

Quand il y a exacte colinéarité $\lambda_p = 0$ on obtient alors une solution des équations normales avec $\mathbf{y}^* = \sum_{j=1}^{p-1} \alpha_j \mathbf{c}_j$.

Si l'on ne retient que k composantes principales en éliminant celles de faibles variances on aura une solution approchée en projetant $\mathbf{y}$ sur un sous-espace de W.

Il suffit alors d'exprimer les $\mathbf{c}_j$ en fonction des variables initiales pour obtenir une formule de régression.

On notera que les composantes principales de forte variance ne sont pas nécessairement les plus explicatives et qu'il vaut mieux les ordonner en fonction de leurs corrélations avec $\mathbf{y}$. Par ailleurs les composantes principales de variance proche de zéro fournissent les relations linéaires approchées existant entre les prédicteurs.

17.5.2 La régression « ridge »

Hoerl et Kennard en 1970 ont proposé de prendre comme estimateur :

$$\mathbf{b}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\,\mathbf{X}'\mathbf{y}$$

où k est une constante positive que l'on ajoute aux termes diagonaux de $\mathbf{X}'\mathbf{X}$ et qui permet d'inverser la matrice sans difficulté numérique.

- **Diminution de l'erreur quadratique.** Il existe des valeurs de k telles que l'erreur quadratique de l'estimation de β est inférieure à celle des moindres carrés au sens où :

$$E\left(\|b_R - \beta\|^2\right) \leq E\left(\|b - \beta\|^2\right) = \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{1}{\lambda_k}$$

En effet l'erreur quadratique est égale à la variance augmentée du carré du biais : dans certaines circonstances un léger biais peut être plus que compensé par une faible variance d'où une erreur quadratique inférieure à la variance de l'estimateur sans biais de variance minimale.

La démonstration se fait aisément pour la régression simple (*cf.* Birkes et Dodge 1993) :

Considérons le modèle $Y = \alpha + \beta x + \varepsilon$ où les x_i sont centrés : dans ces conditions les estimateurs des moindres carrés sont $\hat{\alpha} = \bar{y}$ et $\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$.

Soit $\hat{\beta}_R = c\hat{\beta}$ un estimateur « rétréci » avec $0 < c < 1$.

L'erreur quadratique de l'estimateur rétréci vaut :

$$E((c\hat{\beta} - \beta)^2) = V(c\hat{\beta}) + (E(c\hat{\beta}) - \beta)^2 = c^2V(\hat{\beta}) + (c-1)^2\beta^2 = c^2\frac{\sigma^2}{\sum x_i^2} + (c-1)^2\beta^2$$

En annulant la dérivée par rapport à c, on trouve que le minimum est atteint pour $c = \dfrac{\beta^2}{\beta^2 + \dfrac{\sigma^2}{\Sigma x_i^2}}$ d'où $\hat{\beta}_R = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2 + \dfrac{\sigma^2}{\beta^2}}$ ce qui revient à une régression ridge avec une constante k égale à $k = \dfrac{\sigma^2}{\beta^2}$.

Évidemment β, σ et donc la valeur optimale de k sont inconnus, mais le résultat est prouvé : il existe bien un estimateur rétréci d'erreur quadratique inférieure à la variance de l'estimateur des moindres carrés.

- **Régression à coefficients bornés.** Dans le cas de la régression multiple, on obtient l'estimateur ridge comme solution du problème suivant consistant à trouver des coefficients de régression bornés :

$$\min \|\mathbf{y} - \mathbf{Xb}\|^2 \quad \text{sous} \quad \|\mathbf{b}\|^2 \leq c^2$$

il s'agit donc de régulariser la solution pour éviter des coefficients instables.

Le problème de minimisation sous contrainte est équivalent à : $\min\left(\|\mathbf{y} - \mathbf{Xb}\|^2 + k\|\mathbf{b}\|^2\right)$ avec un multiplicateur de Lagrange k. En annulant la dérivée par rapport à $\mathbf{b}$, on a : $2\mathbf{X}'\mathbf{Xb} - 2\mathbf{X}'\mathbf{y} + 2k\mathbf{b} = 0$ soit $(\mathbf{X}'\mathbf{X} + k\mathbf{I})\mathbf{b} = \mathbf{X}'\mathbf{y}$ d'où le résultat.

- **Régression bayésienne.** Le point de vue bayésien donne également une justification éclairante de la régression ridge : On se donne une distribution *a priori* gaussienne sur β $N(\mathbf{0}\,;\psi^2\,\mathbf{I})$ et on suppose que la loi des $\mathbf{Y}/\boldsymbol{\beta}$ est une gaussienne $N(\mathbf{X}\boldsymbol{\beta}\,;\sigma^2\mathbf{I})$. Un calcul simple montre que la loi *a posteriori* de $\boldsymbol{\beta}/\mathbf{Y}$ est une gaussienne dont la densité est telle que (à une constante près) :

$$\ln(f(\boldsymbol{\beta}/\mathbf{y})) = -\frac{1}{2}\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} - \frac{1}{2}\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{\psi^2}$$

La valeur la plus probable *a posteriori*, qui est ici aussi l'espérance *a posteriori*, est alors : $\hat{\beta} = \left(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{\psi^2}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}$. D'où la signification de k comme un rapport de variances.

La valeur de k restant inconnue, une pratique courante consiste à étudier les variations des coefficients de régression selon k et à choisir une valeur donnant des coefficients stables : $k = 0$ n'étant autre que la solution des moindres carrés, au vu d'un graphique appelé « Trace Ridge ». Compte tenu du caractère subjectif de cette méthode, il est préférable de choisir k par validation croisée : on divise les données en s sous-ensembles disjoints, chaque sous ensemble est prédit à l'aide du regroupement des $s - 1$ autres et on cherche la valeur de k qui miminise la somme des carrés des erreurs.

17.5.3 La régression PLS

Proposée par H. et S. Wold, étudiée par M. Tenenhaus, la régression PLS est proche de la régression sur composantes principales en ce qu'elle effectue une projection sur des combinaisons linéaires des prédicteurs non corrélées entre elles, mais avec la différence essentielle que les composantes PLS sont optimisées pour être prédictives de Y, alors que les composantes principales ne font qu'extraire le maximum de variance des prédicteurs sans tenir compte de Y.

Le critère de détermination des composantes PLS est le critère de Tucker, basé sur la covariance :

$$\max \operatorname{cov}^2(\mathbf{y} ; \mathbf{Xw})$$

Posons $\mathbf{t}_1 = w_{11}\mathbf{x}_1 + w_{12}\mathbf{x}_2 + \cdots + w_{1p}\mathbf{x}_p$ avec $\sum_{j=1}^{p} w_{1j}^2 = 1$

Maximiser la covariance aboutit à un compromis entre maximiser la corrélation entre t_1 et y (régression des moindres carrés ordinaires) et maximiser la variance de $\mathbf{t}_1$ (ACP des prédicteurs) puisque :

$$\operatorname{cov}^2(\mathbf{y} ; \mathbf{Xw}) = r^2(\mathbf{y} ; \mathbf{Xw})\, V(\mathbf{Xw})\, V(\mathbf{y})$$

et que $V(\mathbf{y})$ est fixé.

La solution est élémentaire : les w_{1j} sont proportionnels aux covariances $\operatorname{cov}(\mathbf{y} ; \mathbf{x}_j)$: les coefficients sont donc du même signe que les corrélations simples entre $\mathbf{y}$ et les $\mathbf{x}_j$; il ne peut donc y avoir de signes surprenants.

La régression PLS avec une composante s'écrit alors sous la forme $\mathbf{y} = c_1\mathbf{t}_1 + \mathbf{y}_1$

On obtient ensuite la deuxième composante PLS $\mathbf{t}_2$ en itérant le procédé : on effectue la régression de $\mathbf{y}_1$ sur les résidus des régressions des $\mathbf{x}_j$ avec $\mathbf{t}_1$ puis on écrit $\mathbf{y} = c_1\mathbf{t}_1 + c_2\mathbf{t}_2 + \mathbf{y}_2$ etc.

Le nombre de composantes PLS est en général choisi par validation croisée.

On montre aisément que la première composante PLS est toujours plus corrélée avec $\mathbf{y}$ que la première composante principale :

En effet soit $\mathbf{c}_1$ la première composante principale :

$$\operatorname{cov}(\mathbf{y} ; \mathbf{t}_1) = r(\mathbf{y} ; \mathbf{t}_1)\sigma(\mathbf{t}_1)\sigma(\mathbf{y}) \geq \operatorname{cov}(\mathbf{y} ; \mathbf{c}_1) = r(\mathbf{y} ; \mathbf{c}_1)\sigma(\mathbf{c}_1)\sigma(\mathbf{y})$$
$$\text{donc } r(\mathbf{y} ; \mathbf{t}_1)\sigma(\mathbf{t}_1) \geq r(\mathbf{y} ; \mathbf{c}_1)\sigma(\mathbf{c}_1)$$

comme c_1 est la première composante principale, sa variance est maximale :

$$\sigma(\mathbf{c}_1) \geq \sigma(\mathbf{t}_1) \quad \text{d'où} \quad r(\mathbf{y} ; \mathbf{t}_1) \geq r(\mathbf{y} ; \mathbf{c}_1)$$

La propriété reste vraie pour plus d'une composante, c'est à dire que la régression PLS avec k composantes est toujours meilleure que la régression sur les k premières composantes principales mais la démonstration est difficile (De Jong, 1993).

Un des grands avantages de la régression PLS réside dans la simplicité de son algorithme qui ne nécessite ni inversion, ni diagonalisation de matrices, mais seulement une succession de régressions simples, autrement dit des calculs de produits scalaires. On peut donc traiter de très grands ensembles de données.

L'expérience montre que la régression PLS donne en pratique d'excellentes prévisions, même dans le cas d'un petit nombre d'observations et d'un grand nombre de variables.

La régression dite PLS2 est une alternative à l'analyse canonique lorsque l'on cherche à expliquer simultanément plusieurs réponses Y. Le critère de Tucker s'écrit alors :

$$\text{Max } \text{cov}^2(\mathbf{Y_v} ; \mathbf{X_w}) = r^2(\mathbf{Y_v} ; \mathbf{X_w}) \cdot V(\mathbf{Y_v}) \cdot V(\mathbf{X_w})$$

Il est facile de montrer que la première composante PLS des X est vecteur propre de $V_{12}V_{21}$ (voir chapitre 8)

17.6 UN EXEMPLE

On se propose d'étudier la relation existant entre le prix et les variables suivantes : cylindrée, puissance, longueur, largeur, poids et vitesse de pointe de 18 voitures figurant dans le tableau 17.1 :

TABLEAU 17.1

OBS	NOM	CYL	PUIS	LON	LAR	POIDS	VITESSE	FINITION	PRIX
1	ALFASUD-TI-1350	1350	79	393	161	870	165	B	30570
2	AUDI-100-L	1588	85	468	177	1110	160	TB	39990
3	SIMCA-1307-GLS	1294	68	424	168	1050	152	M	29600
4	CITROEN-GS-CLUB	1222	59	412	161	930	151	M	28250
5	FIAT-132-1600GLS	1585	98	439	164	1105	165	B	34900
6	LANCIA-BETA-1300	1297	82	429	169	1080	160	TB	35480
7	PEUGEOT-504	1796	79	449	169	1160	154	B	32300
8	RENAULT-16-TL	1565	55	424	163	1010	140	B	32000
9	RENAULT-30-TS	2664	128	452	173	1320	180	TB	47700
10	TOYOTA-COROLLA	1166	55	399	157	815	140	M	26540
11	ALFETTA-1.66	1570	109	428	162	1060	175	TB	42395
12	PRINCESS-1800-HL	1798	82	445	172	1160	158	B	33990
13	DATSUN-200L	1998	115	469	169	1370	160	TB	43980
14	TAUNUS-2000-GL	1993	98	438	170	1080	167	B	35010
15	RANCHO	1442	80	431	166	1129	144	TB	39450
16	MAZDA-9295	1769	83	440	165	1095	165	M	27900
17	OPEL-REKORD-L	1979	100	459	173	1120	173	B	32700
18	LADA-1300	1294	68	404	161	955	140	M	22100

17.6.1 Résultats de la régression complète

Les calculs ont été effectués avec le logiciel SAS.

17.6.1.1 Analyse de variance de la régression

On trouve (tableau 17.2) :

TABLEAU 17.2

	DDL	SOMME DE CARRES	CARRE MOYEN	F	PROB > F
REGRESSION	6	520591932.37	86765322.06	4.469	0.0156
RESIDUELLE	11	213563857.91	19414896.17		
TOTALE	17	734155790.28			

Comme $F_{5\,\%}(6\,;\,11) = 3.09$ on rejette l'hypothèse $H_0 : \beta_1 = \beta_2 = \ldots = \beta_6 = 0$.

$$R^2 = 0.7091 \qquad \text{et} \qquad \hat{R}^2 = 0.5504$$
$$\hat{\sigma} = 4406.2$$

17.6.1.2 Estimation des paramètres (tableau 17.3)

TABLEAU 17.3

VARIABLE	COEFFICIENT ESTIME	ECART-TYPE	T SI HO : COEFF. = 0	PROB > |T|	FACTEUR D'INFLATION
CONSTANTE	−8239.36	42718.423	−0.193	0.850	0
CYLINDREE	−3.505	5.55	−0.631	0.540	3.772
PUISSANCE	282.168	174.882	1.613	0.134	11.118
LONGUEUR	−15.037	129.747	−0.116	0.909	7.204
LARGEUR	208.693	412.047	0.506	0.622	4.197
POIDS	12.574	24.622	0.511	0.619	9.957
VITESSE	−111.114	222.256	−0.500	0.627	6.375

On constate qu'au seuil 10 % aucun des coefficients n'est significativement différent de 0 et que certains sont négatifs alors que les corrélations entre le prix et les prédicteurs pris isolément sont significativement positifs. Ce phénomène est dû à la forte dépendance entre les prédicteurs (voir également les valeurs du facteur d'inflation de la variance).

Ci-après tableau 17.4 la matrice de corrélation entre les 7 variables :

TABLEAU 17.4

	CYL	PUIS	LON	LAR	POIDS	VITESSE	PRIX
CYL	1.00000	0.79663	0.70146	0.62976	0.78895	0.66493	0.63858
PUIS	0.79663	1.00000	0.64136	0.52083	0.76529	0.84438	0.79870
LON	0.70146	0.64136	1.00000	0.84927	0.86809	0.47593	0.64376
LAR	0.62976	0.52083	0.84927	1.00000	0.71687	0.47295	0.54665
POIDS	0.78895	0.76529	0.86809	0.71687	1.00000	0.47760	0.75329
VITESSE	0.66493	0.84438	0.47593	0.47295	0.47760	1.00000	0.58176
PRIX	0.63858	0.79870	0.64376	0.54665	0.75329	0.58176	1.00000

17.6.1.3 Étude des résidus et de l'influence des observations

Le tableau 17.5 contient les informations essentielles.

TABLEAU 17.5

		PRIX	PRIX ESTIME	E-TYPE PREDICT	LIMITE INF 95 %	LIMITE SUP 95 %
1	ALFASUD-	30570.0	29616.1	2914.0	17989.0	41243.2
2	AUDI-100	39990.0	36259.7	3572.5	23774.4	48744.9
3	SIMCA-13	29600.0	31411.1	2486.0	20276.0	42546.3
4	CITROEN-	28250.0	26445.8	2259.2	15547.2	37344.3
5	FIAT-132	34900.0	37043.0	2160.8	26241.5	47844.5
6	LANCIA-B	35480.0	34972.8	2707.1	23590.6	46355.0
7	PEUGEOT-	32300.0	33749.1	1945.4	23147.9	44350.4
8	RENAULT-	32000.0	26580.0	2760.8	15135.4	38024.5
9	RENAULT-	47700.0	44445.6	3683.5	31805.1	57086.0
10	TOYOTA-C	26540.0	24650.2	3039.9	12868.0	36432.5
11	ALFETTA-	42395.0	38270.5	3006.8	26529.5	50011.5
12	PRINCESS	33990.0	34830.4	2018.2	24163.4	45497.4
13	DATSUN-2	43980.0	44872.4	3343.6	32698.2	57046.6
14	TAUNUS-2	35010.0	36343.5	2320.9	25382.3	47304.7
15	RANCHO	39450.0	35638.1	2453.2	24538.2	46737.9
16	MAZDA-92	27900.0	32233.4	2726.5	20828.8	43638.0
17	OPEL-REK	32700.0	37103.5	2535.7	25914.1	48292.9
18	LADA-130	22100.0	30389.8	2755.1	18952.0	41827.6
		RESIDU	E-TYPE DU RESID	RESIDU STUDENT.	DISTANCE DE COOK	H
1	ALFASUD-	953.8913	3305.1	0.2886	0.009	0.4374
2	AUDI-100	3730.3	2579.2	1.4463	0.573	0.6574
3	SIMCA-13	−1811.1	3637.9	−0.49785	0.017	0.3183
4	CITROEN-	1804.2	3783.0	0.4769	0.012	0.2629
5	FIAT-132	−2143	3840.0	−.558071	0.014	0.2405
6	LANCIA-B	507.1657	3476.6	0.1459	0.002	0.3775
7	PEUGEOT-	−1449.1	3953.5	−.366544	0.005	0.1949
8	RENAULT-	5420.0	3434.1	1.5783	0.230	0.3926
9	RENAULT-	3254.4	2418.0	1.3459	0.600	0.6988
10	TOYOTA-C	1889.8	3189.6	0.5925	0.046	0.4760
11	ALFETTA-	4124.5	3220.8	1.2806	0.204	0.4657
12	PRINCESS	−840.42	3916.9	−.214564	0.002	0.2098
13	DATSUN-2	−892.42	2869.7	−.310978	0.019	0.5758
14	TAUNUS-2	−1333.5	3745.4	−.356029	0.007	0.2774
15	RANCHO	3811.9	3660.1	1.0415	0.070	0.3100
16	MAZDA-92	−4333.4	3461.4	−1.2519	0.139	0.3829
17	OPEL-REK	−4403.5	3603.5	−1.222	0.106	0.3312
18	LADA-130	−8289.8	3438.7	−2.4108	0.533	0.3910

$Press = 732\,726\,946$ et $\sqrt{\frac{Press}{n}} = 6380.21$

Seul le véhicule n° 18 (le moins cher) présente un résidu studentisé anormalement grand, mais semble avoir une influence normale $\left(\text{le } h_i \text{ moyen vaut } 0.39 = \frac{p+1}{n}\right)$.

Par contre, le véhicule n° 9 (le plus puissant et le plus cher) semble contribuer fortement à la détermination des paramètres.

17.6.2 Recherche d'un modèle restreint

Avec 6 prédicteurs, il y avait 63 modèles possibles. Nous donnons ici les meilleurs modèles à 1, 2, 3, 4, 5, 6 variables (tableau 17.6) :

Tableau 17.6

k	Modèle	R^2	$\hat{\sigma}$
1	Puis	0.638	4076.0
2	Puis. Poids	0.686	3916.4
3	Cyl. Puis. Poids	0.699	3974.4
4	Cyl. Puis. Larg. Poids	0.702	4103.7
5	Cyl. Puis. Larg. Poids Vitesse	0.709	4221.2
6	Complet	0.709	4406.2

On constate que le meilleur modèle au sens de $\hat{\sigma}$ est celui à deux variables (Puissance et Poids) qui fournira les prévisions les plus précises.

Les meilleurs modèles étant emboîtés les diverses techniques de sélection pas à pas donnent ici les mêmes résultats et conduisent au même choix.

Nous reproduisons ci-dessous des sorties de la procédure *Stepwise* du logiciel SAS.

SLENTRY et SLSTAY sont les seuils de signification des tests F d'admission et d'élimination des variables.

La quantité Type II SS représente la perte de somme des carrés expliquée encourue en éliminant la variable correspondante (tableau 17.7).

Le tableau 17.8 montre une amélioration très nette des prévisions en n'utilisant que deux variables au lieu de 6.

TABLEAU 17.7

STEPWISE REGRESSION PROCEDURE FOR DEPENDENT VARIABLE PRIX

NOTE: SLENTRY AND SLSTAY HAVE BEEN SET TO .15 FOR THE STEPWISE TECHNIQUE.

STEP 1 VARIABLE PUIS ENTERED R SQUARE = 0.63792233 C(P) = −0.30837792

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	1	468334369.05604458	468334369.05604458	28.19	0.0001
ERROR	16	265821421.22173311	16613838.82635832		
TOTAL	17	734155790.27777768			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	12363.65292131				
PUIS	257.58978819	48.51607106	468334369.05604458	28.19	0.0001

BOUNDS ON CONDITION NUMBER : 1, 1

STEP 2 VARIABLE POIDS ENTERED R SQUARE = 0.68662695 C(P) = −0.15009700

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	2	504091153.79100612	252045576.89550306	16.43	0.0002
ERROR	15	230064636.48677157	15337642.43245144		
TOTAL	17	734155790.27777768			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	1775.60120140				
PUIS	172.96722456	72.41999846	87492176.38742225	5.70	0.0305
POIDS	16.45116100	10.77448763	35756784.73496154	2.33	0.1476

BOUNDS ON CONDITION NUMBER : 2.413555, 9.654219

NO OTHER VARIABLES MET THE 0.1500 SIGNIFICANCE LEVEL FOR ENTRY INTO THE MODEL.

TABLEAU 17.8

OBS	ACTUAL	PREDICT VALUE	LOWER 95 % PREDICT	UPPER 95 % PREDICT	RESIDUAL	STD ERR RESIDUAL	STUDENT RESIDUAL	COOK'S D
1	30570.0	29752.5	20216.1	39289.0	817.4780	3264.5	0.2504	0.009
2	39990.0	34738.6	26136.2	43341.0	5251.4	3792.9	1.3845	0.042
3	29600.0	30811.1	21981.3	39640.9	−1211.1	3676.1	−.329448	0.005
4	28250.0	27280.2	18325.9	36234.6	969.7528	3609.2	0.2687	0.004
5	34900.0	36904.9	28171.0	45638.9	−2004.9	3726.2	−.538066	0.010
6	35480.0	33726.2	25139.5	42312.8	1753.8	3800.8	0.4614	0.004
7	32300.0	34523.4	25565.3	43481.4	−2223.4	3607.2	−.616371	0.023
8	32000.0	27904.5	18637.2	37171.7	4095.5	3430.9	1.1937	0.144
9	47700.0	45630.9	36023.3	55238.5	2069.1	3218.3	0.6429	0.066
10	26540.0	24696.5	15275.0	34118.0	1843.5	3337.1	0.5524	0.038
11	42395.0	38067.3	28559.2	47575.3	4327.7	3282.7	1.3183	0.245
12	33990.0	35042.3	26191.4	43893.1	−1052.3	3665.0	−.287114	0.004
13	43980.0	44204.9	34599.8	53810.0	−224.92	3219.9	−.069854	0.001
14	35010.0	36493.6	27676.7	45310.5	−1483.6	3682.9	−.402845	0.007
15	39450.0	34186.3	25431.9	42940.7	5263.7	3715.6	1.4166	0.074
16	27900.0	34145.9	25549.9	42741.9	−6245.9	3796.1	−1.6453	0.058
17	32700.0	37497.6	28742.6	46252.6	−4797.6	3715.3	−1.2913	0.062
18	22100.0	29248.2	20470.3	38026.1	−7148.2	3703.4	−1.9302	0.147

La statistique *Press* vaut maintenant 308 496 438 (elle est donc réduite dans un rapport de 2.4) et $\sqrt{\frac{Press}{n}} = 4139.9$.

Si l'on souhaite une formule contenant les 6 prédicteurs, on a le choix entre la régression ridge, la régression sur composantes principales et la régression PLS.

- Régression ridge

Le tableau 17.9 et la figure 17.8 donnent l'évolution des coefficients de régression en fonction du paramètre k La valeur $k = 0.25$ semble convenir et donne un RMSE de 4706.

TABLEAU 17.9 Coefficients de Régression

K	cylindrée	puissance	longueur	largeur	poids	vitesse
0.0	−3.50518	282.169	−15.0377	208.694	12.5747	−111.114
0.05	−2.18019	197.405	2.76652	108.987	15.2924	−26.2437
0.1	−1.30002	163.095	12.6414	78.4137	14.811	3.09658
0.15	−0.693863	142.962	18.2783	67.2553	14.0478	18.3139
0.2	−0.255884	129.251	21.7857	63.497	13.3264	27.6233
0.25	0.0724271	119.112	24.1123	62.9383	12.6918	33.8481
0.3	0.325527	111.21	25.727	63.8295	12.1402	38.2416
0.35	0.524946	104.817	26.8832	65.3631	11.6592	41.4531
0.4	0.684805	99.501	27.7286	67.1422	11.2366	43.8555
0.45	0.814737	94.9847	28.3541	68.9656	10.8621	45.6797
0.5	0.921532	91.0816	28.819	70.7303	10.5273	47.0767

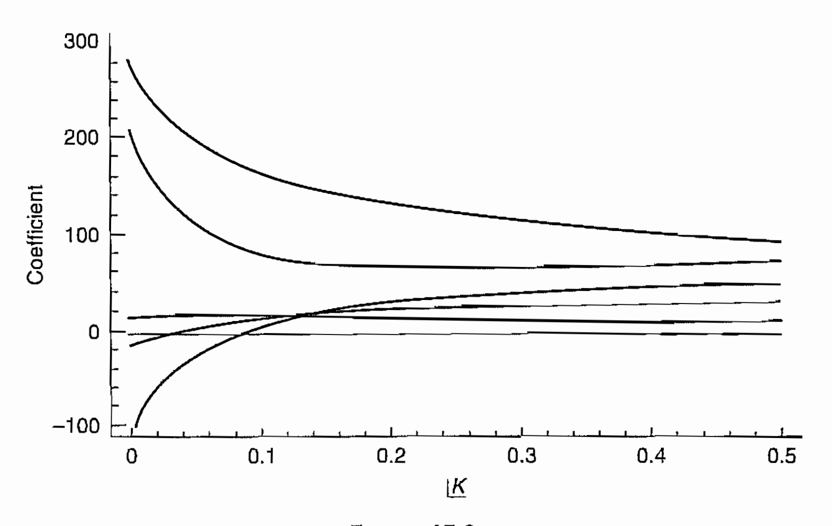


FIGURE 17.8

• Régression sur composantes principales

La régression sur composantes principales fournit les résultats suivants ordonnés selon le nombre de composantes principales conservées :

dim	RMSE	Intercept	CYL	PUIS	LON	LAR	POIDS	VITESSE
1	4301.68	−43286.46	2.74369	49.978	46.0278	175.804	7.5893	71.383
2	4401.15	−34893.04	2.94823	62.544	34.5556	124.103	6.4980	102.827
3	4451.25	−5360.02	4.31052	75.618	30.1484	−39.880	11.5931	45.222
4	4296.24	−5829.58	−2.62099	131.959	70.7514	−167.635	18.6615	64.667
5	4294.23	−9856.87	−4.01533	181.544	−42.9173	141.908	26.3105	11.216
6	4406.23	−8239.36	−3.50518	282.169	−15.0377	208.694	12.5747	−111.114

La solution en dimension 6 est celle des moindres carrés ordinaires. La meilleure formule est sans conteste celle obtenue avec une seule composante principale qui donne un RMSE de 4301.68, inférieur à celui de la régression ridge.

Le spectre des valeurs propres de la matrice de corrélation est :

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	4.4209	73.68	73.68	**
2	0.8561	14.27	87.95	****************
3	0.3731	6.22	94.17	*******
4	0.2139	3.57	97.73	****
5	0.0928	1.55	99.28	**
6	0.0433	0.72	100.00	*

Les coefficients de corrélation entre la variable prix et les 6 composantes principales sont :

	CORRELATIONS VARIABLE-FACTEUR					
	1	2	3	4	5	6
PRIX	−0.77	0.09	−0.13	−0.23	−0.16	−0.10

On remarque que l'ordre des corrélations n'est pas celui des valeurs propres

• Régression PLS

La régression PLS avec une seule composante (c'est ce qu'indique la validation croisée) extrait 73.6 % de la variance de y et 60.8 % de la variance des X. On obtient la formule suivante :

$$\text{PRIX} = -39940.366 + 2.562\text{CYL} + 58.807\text{PUIS} + 43.687\text{LON} + 154.34\text{LAR} + 8.252\text{POIDS} + 71.892\text{VITESSE}$$

Le RMSE est cette fois de 4239, inférieur à celui de la régression sur composantes principales comme le prévoyait la théorie. La régression PLS fournit donc la meilleure formule conservant les 6 variables.

17.7 PRÉDICTEURS QUALITATIFS

17.7.1 Le principe de quantification optimale

Supposons que parmi les prédicteurs on ait q variables qualitatives à $m_1, m_2, \ldots, m_q$ catégories respectivement. On cherchera alors à les transformer en q variables numériques discrètes à $m_1, m_2, \ldots, m_q$ valeurs au plus de sorte que la régression fournisse le R^2 le plus élevé.

On sait que la variable numérique obtenue par quantification d'une variable qualitative est une combinaison linéaire des variables indicatrices des catégories. Il suffit donc de remplacer chaque variable qualitative par l'ensemble des variables indicatrices de ses catégories ; ceci revient à utiliser comme matrice $\mathbf{X}$ de variables explicatives la matrice suivante :

$$\mathbf{X} = \left(\begin{array}{c|c|c|c|c|c} 1 & & & & & \\ 1 & \mathbf{X}_0 & \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_q \\ 1 & & & & & \end{array} \right)$$

où $\mathbf{X}_0$ est le tableau des variables quantitatives et les $\mathbf{X}_i$ $i = 1, 2, \ldots, q$ les tableaux disjonctifs associés aux q variables qualitatives.

Les coefficients de régression associés aux variables indicatrices seront donc les quantifications recherchées.

Cependant une difficulté surgit au moment de résoudre l'équation normale $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$ si $q \geq 1$. En effet, il est facile de s'apercevoir que dans ce cas les colonnes de $\mathbf{X}$ ne sont pas linéairement indépendantes : pour tout tableau disjonctif $\mathbf{X}_i$, la somme des colonnes est égale au vecteur $\mathbf{1}$. Il existe donc q relations linéaires entre les colonnes de $\mathbf{X}$.

Dans ce cas, l'équation normale a une infinité de solutions qui correspondent à des pseudo-inverses différentes de $\mathbf{X}'\mathbf{X}$; toutes ces solutions fournissent d'ailleurs le même vecteur $\mathbf{y}^*$ car la projection de $\mathbf{y}$ sur W est unique, mais les coefficients b_j ne sont pas uniques.

Pour obtenir une estimation unique $\mathbf{b}$ il faut donc imposer q contraintes linéaires sur les codages des variables qualitatives. Les plus simples sont en particulier :

a) Pour chaque variable qualitative une des modalités aura un coefficient b_j nul. Ceci revient en fait à supprimer une colonne dans chaque tableau $\mathbf{X}_i$, ce qui rend la matrice $\mathbf{X}$ de plein rang.

b) Pour chaque variable qualitative la somme des coefficients de $\mathbf{b}$ relatifs à cette variable est nulle. On peut vérifier que ceci revient à supprimer une des colonnes de chaque tableau disjonctif et à remplacer les colonnes restantes par leur différence avec la colonne supprimée.

17.7.2 Retour sur l'analyse de la variance

Lorsque toutes les variables explicatives sont qualitatives la régression multiple correspond à l'analyse de la variance décrite au chapitre 14 en ce sens que l'estimation des effets des niveaux n'est autre que l'estimation des coefficients de régression et que les tests des effets des facteurs sont les tests F de nullité des sous-groupes de coefficients de régression correspondant aux indicatrices d'une variable qualitative.

On vérifiera sans peine que le modèle d'analyse de variance à un facteur correspond à la régression suivante :

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & 0 \\ 1 & 1 & 0 & . & 0 \\ . & & & . & \\ . & & & . & . \\ . & & & . & \\ 1 & 0 & 0 & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ . \\ . \\ . \\ \alpha_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

On impose ici la contrainte $\sum_i n_i \alpha_i = 0$ pour pouvoir identifier les paramètres.

L'analyse de variance à deux facteurs avec interaction correspond à effectuer la régression de $\mathbf{y}$ sur un tableau $\mathbf{X}$ de variables explicatives composé de :

$$\begin{bmatrix} 1 & \overbrace{\quad\quad}^{p} & \overbrace{\quad\quad}^{q} & \overbrace{\quad\quad}^{pq} \\ 1 & & & \\ . & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_{12} \\ . & & & \\ . & & & \\ 1 & & & \end{bmatrix}$$

où $\mathbf{X}_1$ et $\mathbf{X}_2$ sont les tableaux des indicatrices des niveaux des deux facteurs A et B et $\mathbf{X}_{12}$ le tableau des indicatrices d'interaction correspondant aux pq combinaisons des niveaux de A et de B.

On retrouve alors le modèle :

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1\boldsymbol{\alpha} + \mathbf{X}_2\mathbf{b} + \mathbf{X}_{12}\boldsymbol{\gamma} + \mathbf{e}$$

On peut donc utiliser, moyennant les précautions dues au rang de $\mathbf{X}$, un programme de régression multiple pour effectuer de l'analyse de variance même avec des modèles complexes (plusieurs facteurs avec interactions, plans non équilibrés).

17.7.3 Exemple : prix d'une voiture (suite)

On a ajouté aux deux prédicteurs puissance et poids la finition (variable qualitative à trois modalités TB, B, M). Afin d'obtenir une solution on a choisi la contrainte consistant à donner le coefficient 0 à la catégorie TB.

On trouve alors la relation :

$$\texttt{Prix}^* = 23383.6 + 86.96\ \texttt{Puis} + 8.01\ \texttt{Poids} + \begin{bmatrix} -10056.1 \\ -6243.3 \\ 0 \end{bmatrix} \begin{matrix} \texttt{M} \\ \texttt{B} \\ \texttt{TB} \end{matrix}$$

Seuls les différences entre valeurs associées aux catégories ont ici un sens.

Le R^2 est alors de 0.90 et $\hat{\sigma}$ vaut 2320.0.

Le tableau 17.10 montre les résultats de ce modèle à 4 variables explicatives :

Tableau 17.10

	y_i	y_i^*
1	30570.00	30976.30
2	39990.00	39663.33
3	29600.00	27648.39
4	28250.00	25904.76
5	34900.00	34510.48
6	35480.00	39162.20
7	32300.00	33298.60
8	32000.00	30010.28
9	47700.00	45084.43
10	26540.00	24635.99
11	42395.00	41350.06
12	33990.00	33559.50
13	43980.00	44354.30
14	35010.00	34310.28
15	39450.00	39380.66
16	27900.00	29313.20
17	32700.00	34804.52
18	22100.00	26887.63

18 Analyse discriminante et régression logistique

Le but des méthodes de discrimination consiste à prédire une variable qualitative à k catégories à l'aide de p prédicteurs, généralement numériques.

On peut considérer l'analyse discriminante comme une extension du problème de la régression au cas où la variable à expliquer est qualitative; on verra d'ailleurs que dans le cas de deux catégories, on peut se ramener exactement à une régression linéaire multiple.

Les données consistent en n observations réparties en k classes et décrites par p variables explicatives.

On distingue deux aspects en analyse discriminante :

a) ***descriptif*** : chercher quelles sont les combinaisons linéaires de variables qui permettent de séparer le mieux possible les k catégories et donner une représentation graphique (ainsi qu'en analyse factorielle), qui rende compte au mieux de cette séparation;

b) ***décisionnel*** : un nouvel individu se présente pour lequel on connaît les valeurs des prédicteurs. Il s'agit alors de décider dans quelle catégorie il faut l'affecter. C'est un problème de classement (et non de classification, voir chapitre 11)*.

Ces deux aspects correspondent *grosso modo* à la distinction entre méthodes géométriques et méthodes probabilistes faite dans ce chapitre.

Parmi les innombrables applications de l'analyse discriminante citons quelques domaines :

- *aide à la décision en médecine :* à partir de mesures de laboratoire, on cherche une fonction permettant de prédire au mieux le type d'affection d'un malade, ou son évolution probable afin d'orienter le traitement;
- *finance :* prévision du comportement de demandeurs de crédit.

Le terme discrimination est utilisé dans ce chapitre en un sens assez large : nous y incluons la régression logistique afin de mieux la comparer à l'analyse discriminante linéaire.

Le lecteur désireux d'en savoir plus sur l'utilisation de logiciels se reportera avec profit à Nakache et Confais (2003).

---

* Remarque : en anglais « classification » a les deux acceptions.

18.1 MÉTHODES GÉOMÉTRIQUES

Ces méthodes, essentiellement descriptives, ne reposent que sur des notions de distance et ne font pas intervenir d'hypothèses probabilistes.

On supposera vu que les données consistent en n observations de p variables numériques, appartenant à k classes.

18.1.1 Variances interclasse et intraclasse

Les n individus $\mathbf{e}_i$ de l'échantillon constituent un nuage E, de $\mathbb{R}^p$ partagé en k sous-nuages $E_1, E_2, \ldots, E_k$ de centres de gravité $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$, de matrices de variances $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_k$ (fig. 18.1).

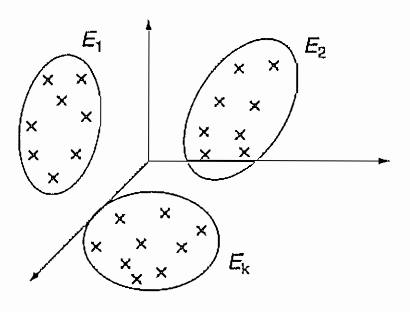


FIGURE 18.1

Soit $\mathbf{g}$ le centre de gravité et $\mathbf{V}$ la matrice de variance de E tout entier. Si les n individus $\mathbf{e}_i$ sont affectés des poids $p_1, p_2, \ldots, p_n$, les poids $q_1, q_2, \ldots, q_k$ de chaque sous-nuage sont alors :

$$q_j = \sum_{\mathbf{e}_i \in E_j} p_i$$

On a :

$$\mathbf{g}_j = \frac{1}{q_j} \sum_i p_i \mathbf{e}_i \qquad \text{pour } \mathbf{e}_i \in E_j$$

$$\mathbf{g} = \sum_{j=1}^{k} q_j \mathbf{g}_j \qquad \text{et} \qquad \mathbf{V}_j = \frac{1}{q_j} \sum_{\mathbf{e}_i \in E_j} p_i (\mathbf{e}_i - \mathbf{g}_j)(\mathbf{e}_i - \mathbf{g}_j)'$$

Appelons matrice de variance interclasse, la matrice de variance $\mathbf{B}$ des k centres de gravité affectés des poids q_j :

$$\boxed{\mathbf{B} = \sum_{j=1}^{k} q_j (\mathbf{g}_j - \mathbf{g})(\mathbf{g}_j - \mathbf{g})'}$$

et matrice de variance intraclasse **W** la moyenne des matrices $\mathbf{V}_j$:

$$\boxed{\mathbf{W} = \sum_{j=1}^{k} q_j \mathbf{V}_j}$$

En règle générale, **W** est inversible tandis que **B** ne l'est pas, car les k centres de gravité sont dans un sous-espace de dimension $k - 1$ de $\mathbb{R}^p$ (si $p > k - 1$ ce qui est généralement le cas), alors que la matrice **B** est de taille p.

On a alors la relation suivante :

$$\boxed{\mathbf{V} = \mathbf{W} + \mathbf{B}}$$

qui se démontre aisément et constitue une généralisation de la relation classique : variance totale = moyenne des variances + variance des moyennes.

Nous supposerons désormais que $\mathbf{g} = \mathbf{0}$, c'est-à-dire que les variables explicatives sont centrées.

Si l'on considère que le tableau de données à étudier se met sous la forme :

$$\begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ n \end{array} \begin{array}{c} \begin{array}{cccc|cccc} 1 & 2 & \dots & k & 1 & 2 & \dots & p \end{array} \\ \left[\begin{array}{cccc|c} 1 & 0 & \dots & 0 & \\ & & & & \\ & & & & \\ & \mathbf{A} & & & \mathbf{X} \\ & & & & \\ & & & & \\ 0 & 0 & \dots & 1 & \end{array}\right] \end{array}$$

où **X** est la matrice des p variables explicatives et **A** le tableau disjonctif associé à la variable qualitative, les k centres de gravité $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_k$ sont les lignes de la matrice $(\mathbf{A'DA})^{-1}(\mathbf{A'DX})$.

A'DA est la matrice diagonale des poids q_j des sous-nuages :

$$\mathbf{A'DA} = \mathbf{D}_q = \begin{bmatrix} q_1 & & & 0 \\ & q_2 & & \\ & & \ddots & \\ 0 & & & q_k \end{bmatrix}$$

La matrice de variance interclasse s'écrit alors, si $\mathbf{g} = \mathbf{0}$:

$$\begin{aligned} \mathbf{B} &= ((\mathbf{A'DA})^{-1}\mathbf{A'DX})'\mathbf{A'DA}((\mathbf{A'DA})^{-1}\mathbf{A'DX}) \\ &= \mathbf{X'DA}(\mathbf{A'DA})^{-1}\mathbf{A'DX} = (\mathbf{X'DA})\mathbf{D}_q^{-1}(\mathbf{A'DX}) \end{aligned}$$

Dans le cas où $p_i = 1/n$ les expressions précédentes se simplifient et en introduisant les effectifs $n_1, n_2, \ldots, n_k$ des k sous-nuages, on a :

$$\mathbf{B} = \frac{1}{n}\sum_j n_j \mathbf{g}_j \mathbf{g}_j' \; ; \; \mathbf{g}_j = \frac{1}{n_j}\sum_{E_j} \mathbf{e}_i ; \; \mathbf{W} = \frac{1}{n}\sum_j n_j \mathbf{V}_j$$

Nous supposerons désormais être dans ce cas.

18.1.2 L'analyse factorielle discriminante (AFD)

18.1.2.1 Les axes et variables discriminantes

L'AFD consiste à rechercher de nouvelles variables (les variables discriminantes) correspondant à des directions de $\mathbb{R}^p$ qui séparent le mieux possible en projection les k groupes d'observations.

L'axe 1 de la figure 18.2 possède un bon pouvoir discriminant tandis que l'axe 2 (qui est l'axe principal usuel) ne permet pas de séparer en projection les deux groupes.

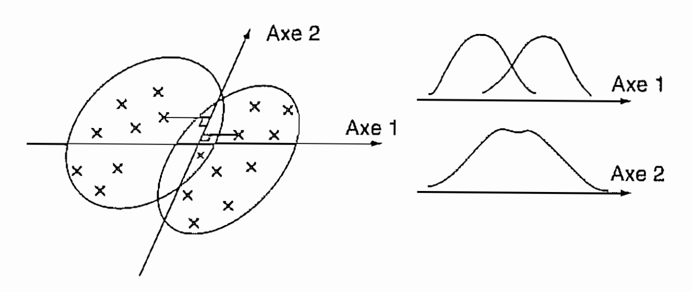


FIGURE 18.2

Supposons $\mathbb{R}^p$ muni d'une métrique $\mathbf{M}$. On notera comme au chapitre 7, $\mathbf{a}$ l'axe discriminant, $\mathbf{u}$ le facteur associé $\mathbf{u} = \mathbf{Ma}$, la variable discriminante sera $\mathbf{Xu}$.

En projection sur l'axe $\mathbf{a}$, les k centres de gravité doivent être aussi séparés que possible, tandis que chaque sous-nuage doit se projeter de manière groupée autour de la projection de son centre de gravité.

En d'autres termes, l'inertie du nuage des $\mathbf{g}_j$ projetés sur $\mathbf{a}$ doit être maximale. La matrice d'inertie du nuage des $\mathbf{g}$ est $\mathbf{MBM}$, l'inertie du nuage projeté sur $\mathbf{a}$ est $\mathbf{a}'\mathbf{MBMa}$ si $\mathbf{a}$ est $\mathbf{M}$-normé à 1.

Il faut aussi qu'en projection sur $\mathbf{a}$, chaque sous-nuage reste bien groupé, donc que $\mathbf{a}'\mathbf{MV}_j\mathbf{Ma}$ soit faible pour $j = 1, 2, \ldots, k$.

On cherchera donc à minimiser la moyenne $\sum_{j=1}^{k} q_j \mathbf{a}'\mathbf{MV}_j\mathbf{Ma}$ soit $\mathbf{a}'\mathbf{MWMa}$.

Or la relation $\mathbf{V} = \mathbf{B} + \mathbf{W}$ entraîne que $\mathbf{MVM} = \mathbf{MBM} + \mathbf{MWM}$, donc que : $\mathbf{a}'\mathbf{MVMa} = \mathbf{a}'\mathbf{MBMa} + \mathbf{a}'\mathbf{MWMa}$.

On prendra alors comme critère, la maximisation du rapport de l'inertie interclasse à l'inertie totale.

Soit :

$$\max_{\mathbf{a}} \frac{\mathbf{a}'\mathbf{MBMa}}{\mathbf{a}'\mathbf{MVMa}}$$

On sait, que ce maximum est atteint si **a** est vecteur propre de $(\mathbf{MVM})^{-1}$ **MBM** associé à sa plus grande valeur propre λ_1 :

$$\mathbf{M}^{-1}\mathbf{V}^{-1}\mathbf{BMa} = \lambda_1\mathbf{a}$$

A l'axe discriminant **a** est alors associé le facteur discriminant **u**, tel que $\mathbf{u} = \mathbf{Ma}$.

On a alors :

$$\boxed{\mathbf{V}^{-1}\mathbf{Bu} = \lambda_1\mathbf{u}}.$$

Les facteurs discriminants, donc les variables discriminantes **Xu**, sont indépendants de la métrique **M**. On choisira par commodité $\mathbf{M} = \mathbf{V}^{-1}$ qui donne $\mathbf{BV}^{-1}\mathbf{a} = \lambda\mathbf{a}$ et $\mathbf{V}^{-1}\mathbf{Bu} = \lambda\mathbf{u}$.

On a toujours $0 \leq \lambda_1 \leq 1$ car λ_1 est la quantité à maximiser.

- $\lambda_1 = 1$ correspond au cas suivant :

En projection sur **a** les dispersions intraclasses sont nulles. Les k nuages sont donc chacun dans un hyperplan orthogonal à **a** (fig. 18.3).

Il y a évidemment discrimination parfaite si les centres de gravité se projettent en des points différents.

- $\lambda_1 = 0$ correspond au cas où le meilleur axe ne permet pas de séparer les centres de gravité $\mathbf{g}_i$, c'est le cas où ils sont confondus.

Les nuages sont donc concentriques et aucune séparation linéaire n'est possible (fig. 18.4).

Il se peut cependant qu'il existe une possibilité de discrimination non linéaire : la distance au centre permet ici de séparer les groupes, mais il s'agit d'une fonction quadratique des variables.

La valeur propre λ est une mesure pessimiste du pouvoir discriminant d'un axe. La figure 18.5 montre qu'on peut discriminer parfaitement car les groupes sont bien séparés malgré $\lambda < 1$.

Le nombre des valeurs propres non nulles, donc d'axes discriminants, est égal à $k - 1$ dans le cas habituel où $n > p > k$ et où les variables ne sont pas liées par des relations linéaires.

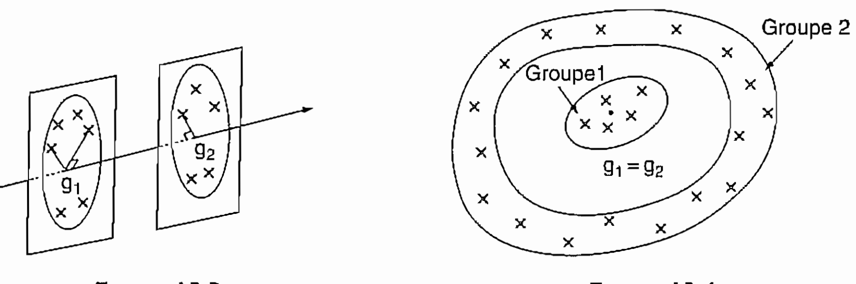


FIGURE 18.3

FIGURE 18.4

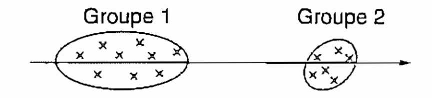


FIGURE 18.5

18.1.2.2 Une analyse en composantes principales (ACP) particulière

D'après les équations précédentes l'analyse factorielle discriminante n'est autre que l'ACP du nuage des k centres de gravité avec la métrique $\mathbf{V}^{-1}$.

On en déduit que les variables discriminantes sont non corrélées 2 à 2.

S'il existe un second axe discriminant, il est possible de représenter le nuage des n observations en projection sur le plan défini par ces deux axes : ce plan est alors celui qui permet le mieux de visualiser la séparation des observations en classes.

Ainsi qu'en ACP, on pourra interpréter les variables discriminantes au moyen d'ur cercle des corrélations.

Nous verrons plus loin que l'analyse factorielle discriminante équivaut aussi à l'ACP des $\mathbf{g}_i$ avec pour métrique $\mathbf{W}^{-1}$.

18.1.2.3 Une analyse canonique particulière

L'analyse discriminante est l'analyse canonique des tableaux A et X.

En effet, l'équation de l'analyse canonique de $\mathbf{A}$ et $\mathbf{X}$ donnant les variables canoniques associées à $\mathbf{X}$ s'écrit :

$$(\mathbf{X'DX})^{-1}\mathbf{X'DA}(\mathbf{A'DA})^{-1}\mathbf{A'DXu} = \lambda\mathbf{u}$$

ce qui est identique à $\mathbf{V}^{-1}\mathbf{Bu} = \lambda\mathbf{u}$ d'après le paragraphe 1. C'est une nouvelle preuve que les variables discriminantes sont non corrélées deux à deux.

Si l'on désigne par $\mathbf{Aa}$ la première variable canonique associée à $\mathbf{A}$ solution de l'autre équation de l'analyse canonique :

$$(\mathbf{A'DA})^{-1}\mathbf{A'DX}(\mathbf{X'DX})^{-1}\mathbf{X'DAa} = \lambda\mathbf{a}$$

normée de telle sorte que sa projection sur le sous-espace de $\mathbb{R}^n$ engendré par les p variables explicatives soit identique à $\mathbf{Xu}$, on peut présenter l'analyse discriminante comme la recherche du codage de la variable qualitative qui la rend le plus proche de l'espace engendré par les colonnes de $\mathbf{X}$. Si les p variables explicatives sont centrées, alors la variable codée l'est aussi et $\mathbf{u}$ est le vecteur des coefficients de régression de $\mathbf{Aa}$ sur $\mathbf{X}$.

La première valeur propre λ_1 est alors le carré du coefficient de corrélation multiple.

L'analyse discriminante est donc bien une généralisation de la régression multiple au cas où la variable à expliquer est qualitative.

La figure 18.6 dans $\mathbb{R}^n$ montre l'identité entre les deux conceptions de l'analyse discriminante : analyse canonique d'une part et maximisation de la variance interclasse par rapport à la variance totale d'autre part.

W_X est l'espace engendré par les colonnes de $\mathbf{X}$; W_A est l'espace engendré par les indicatrices de la variable à expliquer.

Si l'on projette **D**-orthogonalement la variable discriminant ξ sur W_A en **Aa**, le théorème de Pythagore s'écrit :

$$\|\xi\|^2 = \|\mathbf{Aa}\|^2 + \|\mathbf{Aa} - \xi\|^2$$

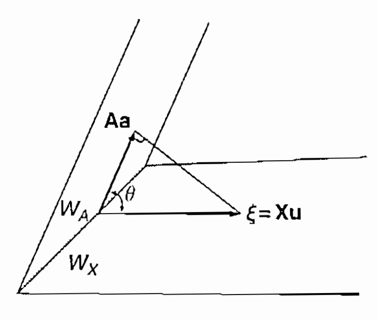


FIGURE 18.6

Variance totale de ξ = variance interclasse + variance intraclasse.

La maximisation du rapport $\dfrac{\text{variance interclasse}}{\text{variance totale}}$ n'est autre que la maximisation de $\cos^2\theta$ où θ est l'angle formé par **Aa** et ξ, ce qui est bien le critère de l'analyse canonique.

On appelle d'ailleurs cette méthode analyse discriminante canonique chez les auteurs anglophones.

18.1.2.4 Analyse de variance et métrique W^{-1}

Si il n'y avait qu'une seule variable explicative on mesurerait l'efficacité de son pouvoir séparateur sur la variable de groupe au moyen d'une analyse de variance ordinaire à un facteur. La statistique F valant alors $\dfrac{\text{variance inter}/k-1}{\text{variance intra}/n-k}$.

Comme il y a p variables on peut rechercher la combinaison linéaire définie par des coefficients **u** donnant la valeur maximale pour la statistique de test ce qui revient à maximiser :

$$\frac{\mathbf{u'Bu}}{\mathbf{u'Wu}}$$

La solution est donnée par l'équation :

$$\mathbf{W}^{-1}\mathbf{Bu} = \mu\mathbf{u} \qquad \text{avec } \mu \text{ maximal}$$

Les vecteurs propres de $\mathbf{W}^{-1}\mathbf{B}$ sont les mêmes que ceux de $\mathbf{V}^{-1}\mathbf{B}$ avec $\mu = \dfrac{\lambda}{1-\lambda}$.

En effet, $\mathbf{Bu} = \lambda\mathbf{Vu}$ est équivalent à :

$$\mathbf{Bu} = \lambda(\mathbf{W} + \mathbf{B})\mathbf{u} \qquad \text{soit} \qquad (1-\lambda)\mathbf{Bu} = \lambda\mathbf{Wu}$$

d'où : $$\boxed{\mathbf{W}^{-1}\mathbf{B}\mathbf{u} = \frac{\lambda \mathbf{u}}{1-\lambda}}.$$

Si $0 \leq \lambda \leq 1$ on a en revanche $0 \leq \mu \leq \infty$ et $\lambda = \dfrac{\mu}{1+\mu}$.

L'utilisation de $\mathbf{V}^{-1}$ ou de $\mathbf{W}^{-1}$ comme métrique est donc indifférent. La métrique $\mathbf{W}^{-1}$ est appelée « métrique de Mahalanobis ».

La convention usuelle dans la plupart des logiciels est d'avoir des variables discriminantes dont la variance intraclasse vaut 1.

On doit donc avoir $\mathbf{u}'\mathbf{W}\mathbf{u} = 1$. Ce qui revient à $\mathbf{u}'\mathbf{B}\mathbf{u} = \dfrac{\lambda}{1-\lambda} = \mu$ et à $\mathbf{u}'\mathbf{V}\mathbf{u} = \dfrac{1}{1-\lambda}$ car $\mathbf{u}'\mathbf{B}\mathbf{u} = \mathbf{u}'\lambda(\mathbf{W}+\mathbf{B})\mathbf{u} = \lambda\mathbf{u}'\mathbf{V}\mathbf{u}$.

18.1.2.5 Un exemple classique : les iris de Fisher

Ce fameux exemple sert de jeu d'essai. Les données concernent trois espèces d'iris (setosa, versicolor, virginica) représentées chacune par 50 individus décrits par 4 variables (longueur et largeur des pétales et sépales).

Il y a donc uniquement deux axes discriminants ce qui permet une représentation plane.

On trouve :

$$\lambda_1 = 0.969872 \qquad \mu_1 = 32.1919$$
$$\lambda_2 = 0.222027 \qquad \mu_2 = 0.2854$$

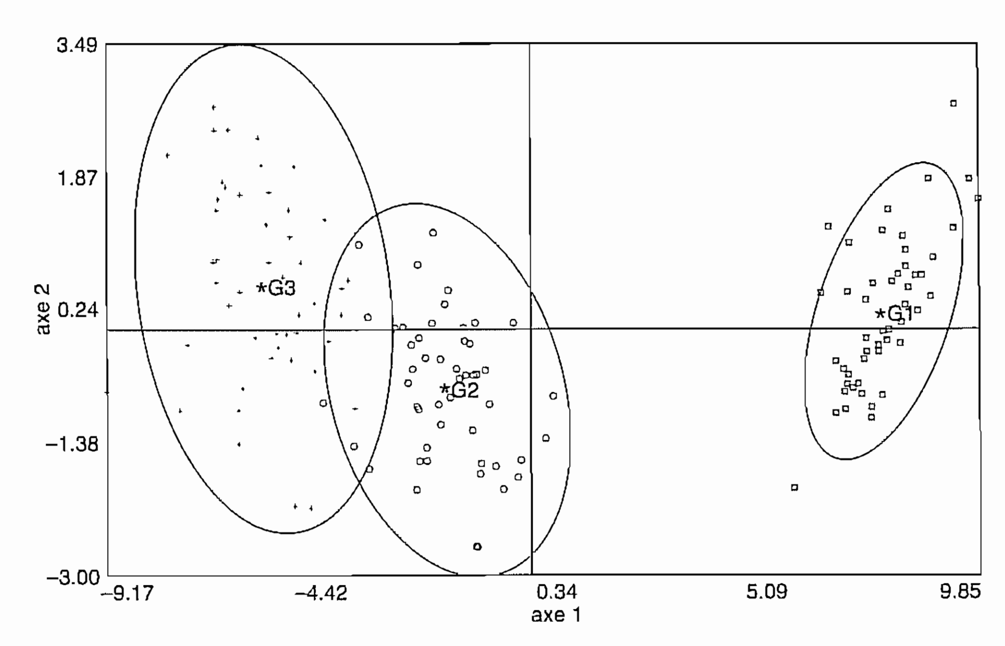


FIGURE 18.7 : Plan discriminant des iris de Fisher

La figure 18.7 montre, avec les ellipses de tolérance à 0.95 (voir 13.6.2), une bonne séparation. L'axe n°2 n'est guère discriminant, mais il est bien utile pour faire un graphique et l'écarter au vu d'un test statistique serait maladroit.

18.1.3 Règles géométriques d'affectation

Ayant trouvé la meilleure représentation de la séparation en k classes des n individus, on peut alors chercher à affecter une observation $\mathbf{e}$ à l'un des groupes.

La règle naturelle consiste à calculer les distances de l'observation à classer à chacun des k centres de gravité et à affecter selon la distance la plus faible. Encore faut-il définir la métrique à utiliser.

18.1.3.1 Règle de Mahalanobis-Fisher

Elle consiste à utiliser la métrique $\mathbf{W}^{-1}$ (ou $\mathbf{V}^{-1}$ ce qui équivalent) :

$$d^2(\mathbf{e}; \mathbf{g}_i) = (\mathbf{e} - \mathbf{g}_i)'\mathbf{W}^{-1}(\mathbf{e} - \mathbf{g}_i)$$

En développant cette quantité on trouve :

$$d^2(\mathbf{e}; \mathbf{g}_i) = \mathbf{e}'\mathbf{W}^{-1}\mathbf{e} + \mathbf{g}_i'\mathbf{W}^{-1}\mathbf{g}_i - 2\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_i$$

Comme $\mathbf{e}'\mathbf{W}^{-1}\mathbf{e}$ ne dépend pas du groupe i, la règle consiste donc à chercher le minimum de $\mathbf{g}'_i\mathbf{W}^{-1}\mathbf{g}_i - 2\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_i$ ou le maximum de $\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_i - (\mathbf{g}'_i\mathbf{W}^{-1}\mathbf{g}_i)/2$.

On voit que cette règle est linéaire par rapport aux coordonnées de $\mathbf{e}$.

Il faut donc calculer pour chaque individu k fonctions linéaires de ses coordonnées et en chercher la valeur maximale.

Illustrons cette règle avec les iris de Fisher : les trois fonctions de classement sont données par le tableau suivant.

Variable	Setosa	Versicolor	Virginica
Constant	−85.20986	−71.75400	−103.26971
SepalLength	2.35442	1.56982	1.24458
SepalWidth	2.35879	0.70725	0.36853
PetalLength	−1.64306	0.52115	1.27665
PetalWidth	−1.73984	0.64342	2.10791

Si l'on applique ces règles aux 150 observations dont on dispose, le tableau suivant (appelé matrice de confusion) donne les résultats de classement : on trouve que les 50 setosa sont

parfaitement classés et que seuls deux versicolor sont attribués à l'espèce virginica, alors qu'un seul virginica est mal classé. Ces résultats semblent excellents, mais sont biaisés en ce sens qu'ils surestiment les performances (voir le paragraphe 18.7.2)

De Espece	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	1	49

Remarquons que l'application de la règle géométrique peut se faire indifféremment dans l'espace $\mathbb{R}^p$ ou dans l'espace factoriel $\mathbb{R}^{k-1}$.

En particulier si $k = 3$, les frontières d'affectation aux groupes sont des hyperplans orthogonaux au plan des trois centres de gravité. On peut lire directement les distances de Mahalanobis à $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$ en utilisant le graphique des deux variables canoniques discriminantes normalisées à 1 (au sens de la variance intraclasse).

18.1.3.2 Insuffisance des règles géométriques

L'utilisation de la règle précédente conduit à des affectations incorrectes lorsque les dispersions des groupes sont très différentes entre elles : rien ne justifie alors l'usage de la même métrique pour les différents groupes.

En effet, si l'on considère la figure 18.8, bien que $\mathbf{e}$ soit plus proche de $\mathbf{g}_1$ que de $\mathbf{g}_2$ au sens habituel il est plus naturel d'affecter $\mathbf{e}$ à la deuxième classe qu'à la première dont le « pouvoir d'attraction » est moindre.

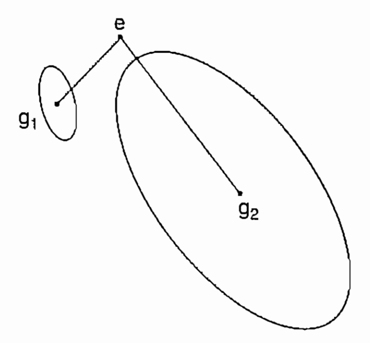


FIGURE 18.8

Diverses solutions utilisant des métriques locales $\mathbf{M}_i$ telles que :

$$d^2(\mathbf{e}; \mathbf{g}_i) = (\mathbf{e} - \mathbf{g}_i)'\mathbf{M}_i\,(\mathbf{e} - \mathbf{g}_i)$$

ont été proposées, la plupart prenant $\mathbf{M}_i$ proportionnel à $\mathbf{V}_i^{-1}$.

La question de l'optimalité d'une règle de décision géométrique ne peut cependant être résolue sans référence à un modèle probabiliste. En effet le problème est de savoir comment cette règle se comportera pour de nouvelles observations ce qui impose de faire des hypothèses distributionnelles sur la répartition dans l'espace de ces nouvelles observations. On atteint donc ici les limites des méthodes descriptives. Nous verrons plus loin dans quelles conditions elles conduisent à des règles optimales.

18.2 FONCTION DE FISHER ET DISTANCE DE MAHALANOBIS POUR DEUX GROUPES

18.2.1 La fonction de Fisher (1936)

Il n'y a donc qu'une seule variable discriminante puisque $k - 1 = 1$.

L'axe discriminant est alors nécessairement la droite reliant les deux centres de gravité $\mathbf{g}_1$ et $\mathbf{g}_2$:

$$\mathbf{a} = (\mathbf{g}_1 - \mathbf{g}_2)$$

La variable discriminante **d** n'obtient en projetant sur **a** selon la métrique $\mathbf{V}^{-1}$ ou $\mathbf{W}^{-1}$ qui tient compte de l' « orientation » des nuages par rapport à la droite des centres (fig. 18.9).

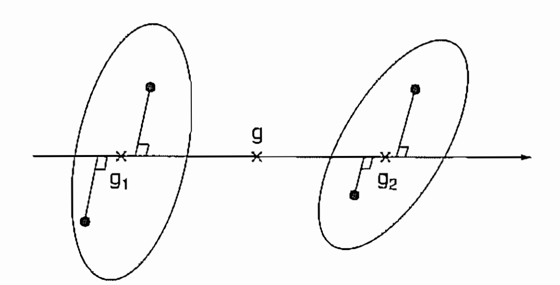


FIGURE 18.9

Le facteur discriminant **u** vaut donc :

$$\mathbf{u} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \qquad \text{ou} \qquad \mathbf{u} = \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

qui lui est proportionnel, (voir plus loin)

$$\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \text{ est la fonction de Fisher}$$

Pour des raisons d'estimation on prend habituellement non pas $\mathbf{W}^{-1}$ mais :

$$\frac{n_1 + n_2 - 2}{n_1 + n_2}\mathbf{W}^{-1}$$

On peut retrouver la démarche de Fisher par le raisonnement suivant :

Cherchons la combinaison linéaire des variables explicatives telles que le carré de la statistique du test T d'égalité des moyennes des deux groupes prenne une valeur maximale :

$$\max \frac{(\bar{y}_1 - \bar{y}_2)^2}{\left(\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)} \qquad \text{où } \mathbf{y} = \mathbf{Xu}$$

en posant $\hat{\Sigma} = \dfrac{n_1 + n_2}{n_1 + n_2 - 2}\mathbf{W}$ ceci revient à maximiser $\dfrac{(\mathbf{u}'(\mathbf{g}_1 - \mathbf{g}_2))^2}{\mathbf{u}'\hat{\Sigma}\mathbf{u}}$. $\mathbf{u}$ est défini à un coefficient multiplicateur près et doit être proportionnel à $\hat{\Sigma}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$.

18.2.2 Application de l'analyse canonique

On peut trouver l'unique valeur propre de $\mathbf{V}^{-1}\mathbf{B}$ en remarquant que pour deux groupes :

$$\mathbf{B} = \frac{n_1 n_2}{n^2}(\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)'$$

En effet : $\mathbf{B} = \dfrac{n_1}{n}\mathbf{g}_1\mathbf{g}_1' + \dfrac{n_2}{n}\mathbf{g}_2\mathbf{g}_2'$; or :

$$\frac{n_1}{n}\mathbf{g}_1 + \frac{n_2}{n}\mathbf{g}_2 = \mathbf{0}$$

On a donc $\mathbf{B} = \dfrac{n_1}{n}\mathbf{g}_1\mathbf{g}_1' - \dfrac{n_1}{n}\mathbf{g}_1\mathbf{g}_2' = \dfrac{n_1}{n}\mathbf{g}_1(\mathbf{g}_1' - \mathbf{g}_2')$ et symétriquement :

$$\mathbf{B} = -\frac{n_2}{n}\mathbf{g}_2(\mathbf{g}_1' - \mathbf{g}_2')$$

donc en moyennant :

$$\mathbf{B} = \frac{n_1 n_2}{n^2}(\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)'$$

On vérifie que $\mathbf{u} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est vecteur propre de $\mathbf{V}^{-1}\mathbf{B}$:

$$\mathbf{V}^{-1}\frac{n_1 n_2}{n^2}(\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = \lambda\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

avec :

$$\lambda = \frac{n_1 n_2}{n^2}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

et :

$$\mu = \frac{\lambda}{1-\lambda} = \frac{n_1 n_2}{n^2}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

μ est donc proportionnel au D_p^2 de Mahalanobis estimé entre les deux groupes (voir chapitre 14 paragr. 14.4.5.2).

On a exactement :

$$\mu = \frac{n_1 n_2}{n(n-2)} D_p^2 \qquad \text{car } D_p^2 = \frac{n-2}{n}(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

On trouve alors :

$$\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = \left(1 + \frac{n_1 n_2}{n(n-2)} D_p^2\right)\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

L'usage de la convention de normalisation $\mathbf{u}'\mathbf{W}\mathbf{u} = 1$ présente l'avantage suivant :

Les coordonnées des deux centres de gravité sur l'axe discriminant ont une différence égale à la distance de Mahalanobis D_p.

En effet $\mathbf{g}_1'\mathbf{u}$ et $\mathbf{g}_2'\mathbf{u}$ sont ces coordonnées où $\mathbf{u}$ est le facteur canonique normalisé. Celui-ci est proportionnel à $\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$, la constante de proportionnalité α étant telle que $\mathbf{u}'\mathbf{W}\mathbf{u} = 1$ soit :

$$[\alpha\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)]'\mathbf{W}[\alpha\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)] = \alpha^2(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

En négligeant la correction par $\frac{n}{n-2}$ (ou en utilisant $\hat{\Sigma}$ à la place de $\mathbf{W}$) il vient $|\alpha| = \frac{1}{D_p}$.

On a donc :

$$|\mathbf{g}_1'\mathbf{u} - \mathbf{g}_2'\mathbf{u}| = |(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{u}| = |\alpha|(\mathbf{g}_1 - \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = \frac{D_p^2}{D_p} = D_p$$

18.2.3 Équivalence avec une régression multiple inhabituelle

L'analyse canonique se réduit ici à une régression multiple puisque après avoir centré, l'espace engendré par les deux indicatrices de la variable des groupes est de dimension 1.

Il suffit donc de définir une variable centrée $\mathbf{y}$ ne prenant que les deux valeurs a et b sur les groupes 1 et 2 respectivement ($n_1 a + n_2 b = 0$).

On obtiendra alors un vecteur des coefficients de régression proportionnel à la fonction de Fisher pour un choix quelconque de a.

Le choix $a = \frac{n}{n_1}$, $b = -\frac{n}{n_2}$ conduit alors à $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$.

On a :

$$R^2 = \frac{D_p^2}{\frac{n(n-2)}{n_1 n_2} + D_p^2}$$

On prendra garde au fait que les hypothèses habituelles de la régression ne sont pas vérifiées bien au contraire : ici $\mathbf{y}$ est non aléatoire et $\mathbf{X}$ l'est. Il ne faudra donc pas utiliser les statistiques usuelles fournies par un programme de régression, en particulier les erreurs standard des coefficients et les niveaux de signification.

Le fait que la fonction de Fisher puisse être obtenue par une régression multiple peu orthodoxe a suscité des controverses et incompréhensions non fondées au profit de la régression logistique (voir 18.6.3 pour une discussion approfondie).

18.2.4 Fonctions de classement et fonction de Fisher

En appliquant la régle du paragraphe 18.1.3.1 au cas de deux groupes on décidera d'affecter au groupe 1 si :

$$\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_1 - \frac{1}{2}(\mathbf{g}_1'\mathbf{W}^{-1}\mathbf{g}_1) > \mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_2 - \frac{1}{2}\mathbf{g}_2'\mathbf{W}^{-1}\mathbf{g}_1$$

soit :

$$\boxed{\mathbf{e}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) > \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)}$$

Comme $\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est la fonction de Fisher, la règle consiste donc à affecter au groupe 1 si la valeur de la fonction discriminante est supérieure au seuil :

$$\frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

Lorsque les deux groupes sont de même effectif $\mathbf{g}_1 + \mathbf{g}_2 = \mathbf{0}$; on affecte au groupe 1 si la fonction $\mathbf{e}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est positive.

18.2.5 Exemple « infarctus »

Les données du tableau 18.1 (communiquées par J.-P. Nakache) concernent 101 victimes d'infarctus du myocarde (51 décèderont, 50 survivront) sur lesquels ont été mesurées à leur admission dans un service de cardiologie 7 variables (fréquence cardiaque, index cardiaque, index systolique, pression diastolique, pression artérielle pulmonaire, pression ventriculaire, résistance pulmonaire). Le tableau 18.2 donne les statistiques élémentaires par groupe.

Tableau 18.1

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONO
90	1.71	19.0	16	19.5	16.0	912	SURVIE
90	1.68	18.7	24	31.0	14.0	1476	DECES
120	1.40	11.7	23	29.0	8.0	1657	DECES
82	1.79	21.8	14	17.5	10.0	782	SURVIE
80	1.58	19.7	21	28.0	18.5	1418	DECES
80	1.13	14.1	18	23.5	9.0	1664	DECES
94	2.04	21.7	23	27.0	10.0	1059	SURVIE
80	1.19	14.9	16	21.0	16.5	1412	SURVIE
78	2.16	27.7	15	20.5	11.5	759	SURVIE
100	2.28	22.8	16	23.0	4.0	807	SURVIE
90	2.79	31.0	16	25.0	8.0	717	SURVIE
86	2.70	31.4	15	23.0	9.5	681	SURVIE
80	2.61	32.6	8	15.0	1.0	460	SURVIE
61	2.84	47.3	11	17.0	12.0	479	SURVIE
99	3.12	31.8	15	20.0	11.0	513	SURVIE
92	2.47	26.8	12	19.0	11.0	615	SURVIE
96	1.88	19.6	12	19.0	3.0	809	SURVIE
86	1.70	19.8	10	14.0	10.5	659	SURVIE
125	3.37	26.9	18	28.0	6.0	665	SURVIE
80	2.01	25.0	15	20.0	6.0	796	SURVIE
82	3.15	38.4	13	20.0	6.0	508	SURVIE
110	1.66	15.1	23	31.0	6.5	1494	DECES
80	1.50	18.7	13	17.0	12.0	907	DECES
118	1.03	8.7	19	27.0	10.0	2097	DECES
95	1.89	19.9	25	27.0	20.0	1143	DECES
80	1.45	18.1	19	23.0	15.0	1269	DECES
85	1.30	15.1	13	18.0	10.0	1108	DECES
105	1.84	17.5	18	22.0	10.0	957	DECES
122	2.79	22.9	25	36.0	10.0	1032	SURVIE
81	1.77	21.9	18	27.0	11.0	1220	SURVIE
118	2.31	19.6	22	27.0	10.0	935	SURVIE
87	1.20	13.8	34	41.0	20.0	2733	DECES
65	1.19	18.3	15	18.0	13.0	1210	DECES
84	2.15	25.6	27	37.0	10.0	1377	SURVIE
103	0.91	8.8	30	33.5	10.0	2945	DECES
75	2.54	33.9	24	31.0	16.0	976	SURVIE
90	2.08	23.1	20	28.0	6.0	1077	SURVIE
90	1.93	21.4	11	18.0	10.0	746	SURVIE
90	0.95	10.6	20	24.0	6.0	2021	DECES
65	2.38	36.6	16	22.0	12.0	739	SURVIE
95	0.99	10.4	20	27.5	8.0	2222	DECES
95	0.85	8.9	19	22.0	15.5	2071	DECES
86	2.05	23.8	21	28.0	10.0	1093	SURVIE
82	2.02	24.6	16	22.0	14.0	871	SURVIE
70	1.44	20.6	19	26.5	11.0	1472	DECES
92	3.06	33.3	10	15.0	6.0	392	SURVIE
94	1.31	13.9	26	40.0	15.0	2443	DECES
79	1.29	16.3	24	31.0	10.0	1922	DECES
67	1.47	21.9	15	18.0	16.0	980	SURVIE
75	1.21	16.1	19	24.0	4.0	1587	DECES

TABLEAU 18.1 (*suite et fin*)

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONO
80	2.41	30.9	19	24.0	7.0	797	SURVIE
61	3.28	54.0	12	16.0	7.0	390	SURVIE
110	1.24	11.3	22	27.5	11.0	1774	DECES
116	1.85	15.9	33	42.0	13.0	1816	DECES
75	2.00	26.7	16	22.0	5.0	880	SURVIE
92	1.97	21.4	18.0	27.0	3.0	1096	DECES
110	0.96	8.8	15.0	19.0	16.0	1583	SURVIE
95	2.56	26.9	8.0	13.0	3.0	406	SURVIE
75	2.32	30.9	8.0	10.0	6.0	345	SURVIE
80	2.65	33.1	13.0	19.0	9.0	574	SURVIE
102	1.60	15.7	24.0	31.0	16.0	1550	DECES
86	1.67	19.4	18.0	23.0	8.5	1102	SURVIE
60	0.82	13.7	22.0	32.0	13.0	3122	DECES
100	1.76	17.6	23.0	33.0	2.0	1500	SURVIE
80	3.28	41.0	12.0	17.0	2.0	415	SURVIE
108	2.96	27.4	24.0	35.0	6.5	946	SURVIE
92	1.37	14.8	25.0	46.0	11.0	2686	DECES
100	1.38	13.8	20.0	31.0	11.0	1797	DECES
80	2.85	35.6	25.0	32.0	7.0	898	SURVIE
87	2.51	28.8	16.0	24.0	20.0	765	DECES
100	2.31	23.1	8.0	12.0	1.0	416	SURVIE
120	1.18	9.9	25.0	36.0	8.0	2441	DECES
115	1.83	15.9	25.0	30.0	8.0	1311	DECES
101	2.55	25.2	23.2	30.5	9.0	957	SURVIE
92	2.17	23.5	19.0	24.0	3.0	885	SURVIE
87	1.42	16.1	20.0	26.0	10.0	1465	DECES
80	1.59	19.9	13.0	20.5	4.0	1031	SURVIE
88	1.47	16.7	23.0	32.5	10.0	1769	DECES
104	1.23	11.8	27.0	33.0	11.0	2146	DECES
90	1.45	16.1	17.0	24.0	8.5	1324	SURVIE
67	0.85	12.7	26.0	33.0	11.0	3106	DECES
87	2.37	27.2	15.0	22.0	10.0	743	SURVIE
108	2.40	22.2	26.0	31.0	4.0	1033	SURVIE
120	1.91	15.9	18.0	27.0	15.0	1131	DECES
108	1.50	13.9	28.0	43.0	16.0	1813	DECES
86	2.36	27.4	24.0	34.0	8.0	1153	SURVIE
112	1.56	13.9	24.0	29.0	4.0	1487	DECES
80	1.34	17.0	16.0	25.0	16.0	1493	DECES
95	1.65	17.4	20.0	33.0	7.0	1600	DECES
90	2.04	22.7	28.0	41.0	10.0	1608	DECES
90	3.03	33.6	17.0	23.5	7.0	620	SURVIE
94	1.21	12.9	17.0	22.0	3.0	1455	DECES
51	1.34	26.3	11.0	17.0	6.0	1015	DECES
110	1.17	10.6	29.0	35.0	10.5	2393	DECES
96	1.74	18.1	24.0	29.0	6.0	1333	DECES
132	1.31	9.9	23.0	28.0	12.0	1710	DECES
135	0.95	7.0	15.0	20.0	7.0	1684	DECES
105	1.92	18.3	18.0	24.0	3.0	1000	DECES
99	0.83	8.4	23.0	27.0	8.0	2602	DECES
116	0.60	5.2	33.0	38.0	10.0	5067	DECES
112	1.54	13.8	25.0	31.0	8.0	1610	DECES

TABLEAU 18.2

PRONO = DECES

VARIABLE	N	MEAN	STANDARD DEVIATION
FRCAR	51	95.90196078	17.97693511
INCAR	51	1.39470588	0.37619332
INSYS	51	14.99607843	4.63900682
PRDIA	51	21.96078431	5.14183152
PAPUL	51	29.09803922	6.81910523
PVENT	51	10.64705882	4.34429985
REPUL	51	1797.27450980	739.87296419

PRONO = SURVIE

VARIABLE	N	MEAN	STANDARD DEVIATION
FRCAR	50	88.34000000	13.84109527
INCAR	50	2.30580000	0.56055035
INSYS	50	26.75200000	8.08319597
PRDIA	50	16.50400000	5.15304388
PAPUL	50	22.84000000	6.46532352
PVENT	50	8.33000000	4.05398519
REPUL	50	841.38000000	303.68256050

La distance de Mahalanobis au carré vaut :

$$D_7^2 = 4.942 \qquad \text{d'où} \qquad D_7 = 2.223$$

Sous les hypothèses de multinormalité du chapitre 14 paragraphe 14.4.5.2, cette valeur correspond à un $F = 16.476$:

$$\frac{n_1 n_2}{n} \frac{n - p - 1}{p(n - 2)} D_p^2 = F$$

La valeur critique à 1 % pour un $F(7; 93)$ étant de 2.84, le D^2 est significatif d'une différence nette entre les deux groupes.

On trouve $R^2 = \lambda = 0.5576$ et $\mu = 1.2604$.

La variable discriminante s'obtient alors par la combinaison linéaire des 7 variables centrées sur la moyenne générale des deux groupes (tableau 18.3).

TABLEAU 18.3

FRCAR	−0.026445290
INCAR	2.768181397
INSYS	−0.075037835
PRDIA	0.009115031
PAPUL	−0.074211897
PVENT	−0.021086258
REPUL	0.000084078

ou si l'on ne centre pas en ajoutant la constante 1.22816 à la combinaison linéaire précédente des données brutes.

Les coefficients de corrélation linéaires de la variable discriminante avec les 7 variables (les deux groupes confondus) sont indiqués sur le tableau 18.4.

Tableau 18.4

FRCAR	−0.3097
INCAR	0.9303
INSYS	0.8976
PRDIA	−0.6321
PAPUL	−0.5751
PVENT	−0.3592
REPUL	−0.8676

Les moyennes des deux groupes sur la variable discriminante sont :

Décès	−1.1005
Survie	1.1225

On retrouve $D_7 = +\ 1.1005 + 1.1225 = 2.2230$.

En appliquant les régles géométriques de classement le tableau 18.5 donne pour l'exemple des infarctus les deux fonctions suivantes

Tableau 18.5

	DECES	SURVIE
CONSTANT	−91.57481116	−89.97034555
FRCAR	1.53609883	1.47730875
INCAR	−52.09444392	−45.94054613
INSYS	5.44165359	5.27483824
PRDIA	−0.64815662	−0.62789315
PAPUL	0.70738671	0.54240748
PVENT	0.85037707	0.80350057
REPUL	0.00638975	0.00657667

La fonction de Fisher s'obtient par différence entre les deux fonctions de classement (survie – décès). En divisant ensuite les coefficients par la distance de Mahalanobis, on retrouve les coefficients du tableau 18.3.

18.3 LES SVM OU SÉPARATEURS À VASTE MARGE

Lorqu'il n'y a que deux groupes, l'établissement d'une règle linéaire est équivalente à la détermination d'un hyperplan séparateur, ou frontière plane, et réciproquement.

A la fonction de Fisher (figure 18.10a) est associé l'hyperplan médiateur (figure 18.10b) de $\mathbf{g}_1$ et $\mathbf{g}_2$ (au sens de la mètrique $\mathbf{V}^{-1}$ ou $\mathbf{W}^{-1}$).

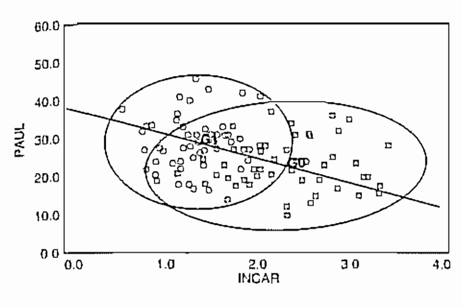


FIGURE 18.10a Axe discriminant

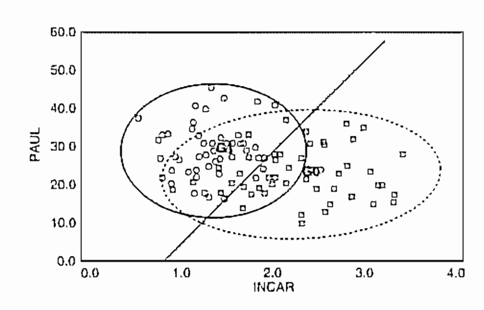


FIGURE 18.10b Frontière

Si l'on note que la fonction de Fisher ne fait que définir une combinaison lnéaire sans terme constant, changer le seuil de classement revient à déplacer l'hyperplan parallèlement à lui-même.

On peut chercher directement une frontière, linéaire ou non, à condition de définir un critère convenable.

18.3.1 L'hyperplan optimal

La recherche directe d'un hyperplan optimal a fait l'objet de nombreux travaux depuis le perceptron de Rosenblatt (1958). On doit à V. Vapnik (1986) d'avoir défini un critère d'optimalité basé sur la « marge », ou largeur d'une zone de part et d'autre de la frontière, et de l'avoir généralisé à des frontières non-linéaires grâce à un changement d'espace.

L'objectif étant de classer, on peut chercher à minimiser le nombre d'observations mal classées, ou points du mauvais côté de la frontière. Ce critère ne suffit cependant pas à déterminer de manière unique un hyperplan séparateur : pour des données linéairement séparables, il y a une infinité de solutions comme le montre la figure 18.11.

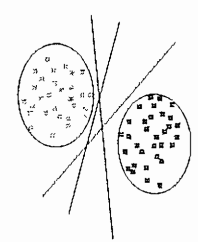


FIGURE 18.11

18.3.1.1 Le cas séparable

Soit C la plus petite distance d'un point à la frontière. Vapnik a proposé que l'hyperplan optimal soit celui qui maximise cette distance, ce qui revient à avoir le plus grand « no man's land » de part et d'autre de la frontière. La marge qui est la largeur du couloir vaut donc $2C$. En dimension 2, on voit sur la figure 18.12 qu'il suffit de trouver les 3 points $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ (appelés points support) pour définir la frontière : on trace la parallèle à $\mathbf{x}_2$, $\mathbf{x}_3$ passant par $\mathbf{x}_1$, puis la droite au milieu.

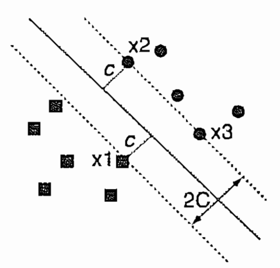


FIGURE 18.12

La solution mathématique est la suivante :

L'hyperplan séparateur a pour équation $\sum_{j=1}^{p} x^j \beta_j + \beta_0 = \mathbf{x}'\boldsymbol{\beta} + \beta_0 = 0$ où les coefficients β sont définis à un facteur multiplicatif près.

La distance d'un point $\mathbf{x}_i$ à l'hyperplan vaut :

$$\frac{|\mathbf{x}_i'\boldsymbol{\beta} + \beta_0|}{\|\boldsymbol{\beta}\|}$$

Notons $y_i = 1$ ou $y_i = -1$ les appartenances aux deux groupes. Pour que les points soient tous du bon côté et à une distance supérieure à C, il faut pour tout i :

$$\frac{1}{\|\boldsymbol{\beta}\|} y_i (\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) \geq C$$

et on doit maximiser C sous ces n contraintes.

Les β étant définis à une constante près, on choisit $\|\boldsymbol{\beta}\| = \frac{1}{C}$. Maximiser C revient à minimiser $\|\boldsymbol{\beta}\|$, d'où le programme quadratique suivant :

$$\begin{cases} \min_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\|^2 \\ y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) \geq 1 \end{cases}$$

Ce problème admet une solution unique, dont les propriétés sont les suivantes. Soient α_i les multiplicateurs de Lagrange associées aux contraintes. En dérivant le lagrangien $L = \|\boldsymbol{\beta}\|^2 - 2\sum_{i=1}^{n} \alpha_i [y_i (\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) - 1]$ (le facteur 2 est introduit par commodité), on trouve $\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$ ainsi que les conditions de Kuhn et Tucker

$$\alpha_i [y_i (\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) - 1] = 0$$

Si $\alpha_i > 0$ alors $y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) = 1$ et le point est sur la marge

Si $y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) > 1$ alors $\alpha_i = 0$

La règle de classement est basée sur le signe de

$$f(\mathbf{x}) = \beta_0 + \sum_{\text{support}} \alpha_i y_i \mathbf{x}_i'\mathbf{x}$$

L'hyperplan optimal ne dépend que des points support où α_i est non nul, situés sur la marge, donc les plus difficiles à classer, ce qui le différencie de l'hyperplan de Fisher : il peut être plus robuste, car il ne dépend pas des points situés loin de la frontière. On verra au chapitre suivant une propriété supplémentaire concernant la généralisation à de futures données.

18.3.1.2 Le cas non-séparable

Dans ce cas certains points seront du mauvais côté de la frontière (figure 18.13), et on va chercher à minimiser l'importance de l'erreur de classement. On introduit alors les variables d'écart ξ^* et ξ. Pour un point mal classé ξ^* est la distance à la marge de sa classe, et on pose $\xi^* = C\xi$.

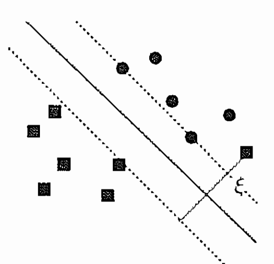


FIGURE 18.13

On modifie alors les contraintes par $y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i$ et on introduit une nouvelle contrainte pour borner l'erreur de classement $\sum \xi_i < cste$. Le problème d'optimisation se transforme en :

$$\begin{cases} \min_{\gamma,\beta,\beta_0} \|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^{n} \xi_i \\ y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i \end{cases}$$

et donne une solution du même type que précédemment:

$$f(\mathbf{x}) = \beta_0 + \sum_{\text{support}} \alpha_i y_i \mathbf{x}_i'\mathbf{x}.$$

Le paramètre γ peut être réglé par l'utilisateur, mais cela est délicat. On préconise une optimisation par validation croisée, ou avec un autre échantillon.

18.3.2 Changement d'espace

Des données non séparables linéairement dans leur espace d'origine E, peuvent le devenir après un changement d'espace Φ(E), en général de dimension plus élevée. A une frontière linéaire dans Φ(E), correspond une frontière non-linéaire dans E.

L'exemple (figure 18.14) suivant est classique avec deux groupes séparées par une parabole : en passant de l'espace $\mathbb{R}^3$ défini par $(1, x_1, x_2)$ à l'espace $\mathbb{R}^6$ défini par $(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$, on obtient une séparation linéaire dans le sous-espace x_2, x_1^2

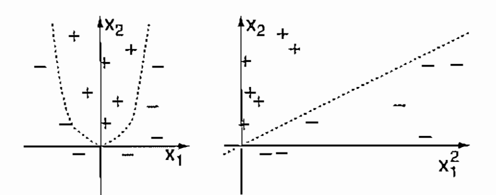


FIGURE 18.14

L'hyperplan optimal dans Φ(E) s'écrit $f(\mathbf{x}) = \beta_0 + \sum_{\text{support}} \alpha_i y_i < \Phi(\mathbf{x}_i) \,;\, \Phi(\mathbf{x}) > = 0$

Son équation ne fait intervenir que les produits scalaires entre points transformés. Comme dans le chapitre 7 avec la kernel-ACP, un choix astucieux du produit scalaire $<\Phi(\mathbf{x}_i) \,;\, \Phi(\mathbf{x}_j)> = K(\mathbf{x}_i \,;\, \mathbf{x}_j)$ évite de calculer explicitement Φ et permet d'effectuer tous les calculs dans E.

Le classifieur écrit alors $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \langle \Phi(\mathbf{x}_i) | \Phi(\mathbf{x}) \rangle + \beta_0$ et la somme n'est à effectuer que sur les points supports.

La capacité prédictive des SVM est élevée. Nous verrons plus loin que le risque de biais de surapprentissage qui paraît élevé, est maitrisé par la maximisation de la marge, à condition de ne pas chercher nécessairement une séparation parfaite dans Φ(E).

Les exemples suivants (figure 18.15) sont obtenus avec le noyau polynomial de degré 3 $K(\mathbf{x}, \mathbf{y}) = (<\mathbf{x}, \mathbf{y}> + c)^3$ et montrent sa flexibilité (logiciel LIB-SVM):

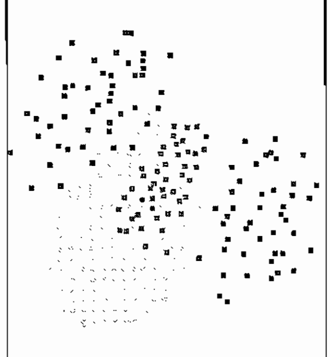
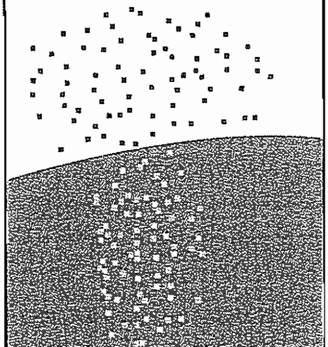
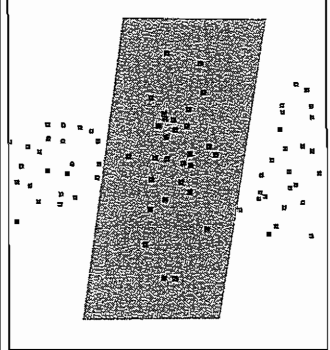




FIGURE 18.15

18.4 DISCRIMINATION SUR VARIABLES QUALITATIVES

Les méthodes précédentes ne s'appliquent pas directement lorsque les prédicteurs ne sont pas numériques, ce qui est pourtant un cas assez courant.

18.4.1 Discriminante sur variables indicatrices

Une solution consiste à transformer (quantifier) les prédicteurs en variables numériques discrètes en attribuant des valeurs (notes ou scores partiels) à leurs modalités. On cherchera des valeurs « optimales » en un certain sens lié aux performances attendues de la discrimination.

On a vu au chapitre 9 paragraphe 9.4.1 et au chapitre 17 paragraphe 17.7.1 que cette transformation revient à remplacer les variables qualitatives par les indicatrices des catégories. Ainsi un problème de discrimination sur p variables qualitatives à $m_1, \ldots, m_p$ catégories revient à une analyse discriminante de $\mathbf{y}$ sur le tableau disjonctif des $m_1 + \ldots + m_p$ indicatrices des prédicteurs $(\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_p)$.

Cette matrice n'étant pas de plein rang, $\mathbf{V}$ et $\mathbf{W}$ ne sont pas inversibles et il y a donc une infinité de solutions équivalentes. On peut alors faire comme pour le modèle linéaire général et éliminer une indicatrice pour chaque prédicteur, ce qui équivaut à lui donner un coefficient nul.

Dans le cas d'une discrimination entre deux classes, la fonction de Fisher calculée sur ces $(m_1 + \ldots + m_p - p)$ indicatrices fournit par ses coefficients la quantification recherchée. Cette quantification rend maximale la distance de Mahalanobis entre les centres de gravité des deux groupes.

18.4.2 Discrimination sur composantes d'une ACM

On sait que l'ensemble des composantes de l'ACM de $\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_p)$ engendre le même espace que $\mathbf{X}$. S'inspirant de la régression sur composantes principales (17.5.1) la méthode Disqual (Saporta, 1975) consiste à effectuer l'analyse discriminante sur une sélection d'axes. En réduisant la dimension de l'espace des prédicteurs et en éliminant des dimensions inutiles, on assure une plus grande robustesse des résultats.

Détaillons les formules dans le cas de deux groupes, qui on le sait est équivalent à une régression après recodage de $\mathbf{y}$.

Notons $\mathbf{z}^j$ les composantes de l'ACM et λ_j les valeurs propres. A l'aide de tests et aussi de l'expertise du statisticien, on éliminera les composantes de faible inertie ainsi que celles ne séparant pas suffisamment les deux groupes : il suffit d'effectuer un test de comparaison de moyennes sur chaque axe. Soit q le nombre de composantes conservées. Comme les composantes sont orthogonales, il est plus simple d'inverser $\mathbf{V}$, qui est diagonale, que $\mathbf{W}$. La fonction de Fisher étant définie à un coefficient multiplicatif près, c'est sans importance.

La variable « score » $\mathbf{s}$ qui donne la valeur de la fonction de Fisher pour chaque observation s'écrit alors $\mathbf{s} = \sum_{j=1}^{q} u_j \mathbf{z}^j$ avec $\mathbf{u} = \begin{pmatrix} \cdot \\ u_j \\ \cdot \end{pmatrix} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = \begin{pmatrix} \cdot \\ \dfrac{\bar{z}_1^j - \bar{z}_2^j}{\lambda_j} \\ \cdot \end{pmatrix}$.

Grâce aux formules de transition de l'ACM (paragraphe 10.1.3.3), il n'est pas nécessaire d'avoir à calculer pour chaque observation ses coordonnées sur les axes factoriels : il suffit d'effectuer la combinaison linéaire avec les mêmes coefficients u_j des coordonnées de ses catégories.

En effet à un facteur multiplicatif près, on a $\mathbf{z}^j = \mathbf{X}\mathbf{a}^j$ où $\mathbf{a}^j$ est le vecteur des coordonnées des $m_1 + \ldots + m_p$ modalités sur l' axe n° j, d'où :

$$\mathbf{s} = \sum_{j=1}^{q} u_j \mathbf{X}\mathbf{a}^j = \mathbf{X} \underbrace{\sum_{j=1}^{q} u_j \mathbf{a}^j}_{\text{grille de score}}$$

Le score s'exprime alors directement comme combinaison linéaire des indicatrices des modalités : pour chaque individu, il suffit d'additionner les scores partiels des modalités qu'il prend. La formule ne comporte pas de terme constant : en pratique ce terme qui correspond au seuil de décision pour classer dans un groupe sera déterminé au vu des erreurs de classement.

18.4.3 Un exemple de « credit scoring »[1]

Les données analysées, provenant du logiciel SPAD, sont relatives à 468 clients d'une banque. On veut prédire la qualité du client (« bon » ou « mauvais ») à partir de 6 caractéristiques qualitatives (voir les résultats pour le détail) totalisant 21 modalités. Il y a donc 15 axes .

L'ACM avec la qualité client en variable supplémentaire montre un bon pouvoir prédictif : valeurs-test élevées pour la variable supplémentaire sur les deux premiers axes.

TABLEAU 18.6

VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	0.3401	13.60	13.60	**********************************
2	0.2466	9.87	23.47	**************************
3	0.2108	8.43	31.90	***********************
4	0.1948	7.79	39.69	*********************
5	0.1843	7.37	47.06	********************
6	0.1758	7.03	54.10	*******************
7	0.1700	6.80	60.90	******************
8	0.1597	6.39	67.28	*****************
9	0.1495	5.98	73.26	****************
10	0.1375	5.50	78.76	***************
11	0.1282	5.13	83.89	*************
12	0.1137	4.55	88.44	************
13	0.1092	4.37	92.81	***********
14	0.1022	4.09	96.90	**********
15	0.0775	3.10	100.00	*****

Dans la figure 18.16 les tailles des points sont proportionnelles aux effectifs des modalités : on identifie facilement les modalités proches des catégories de client, mais d'autres axes vont se révéler nécessaires.

1 Le lecteur intéressé par les applications au domaine financier se reportera utilement à Bardos (2001) et Tufféry (2005).

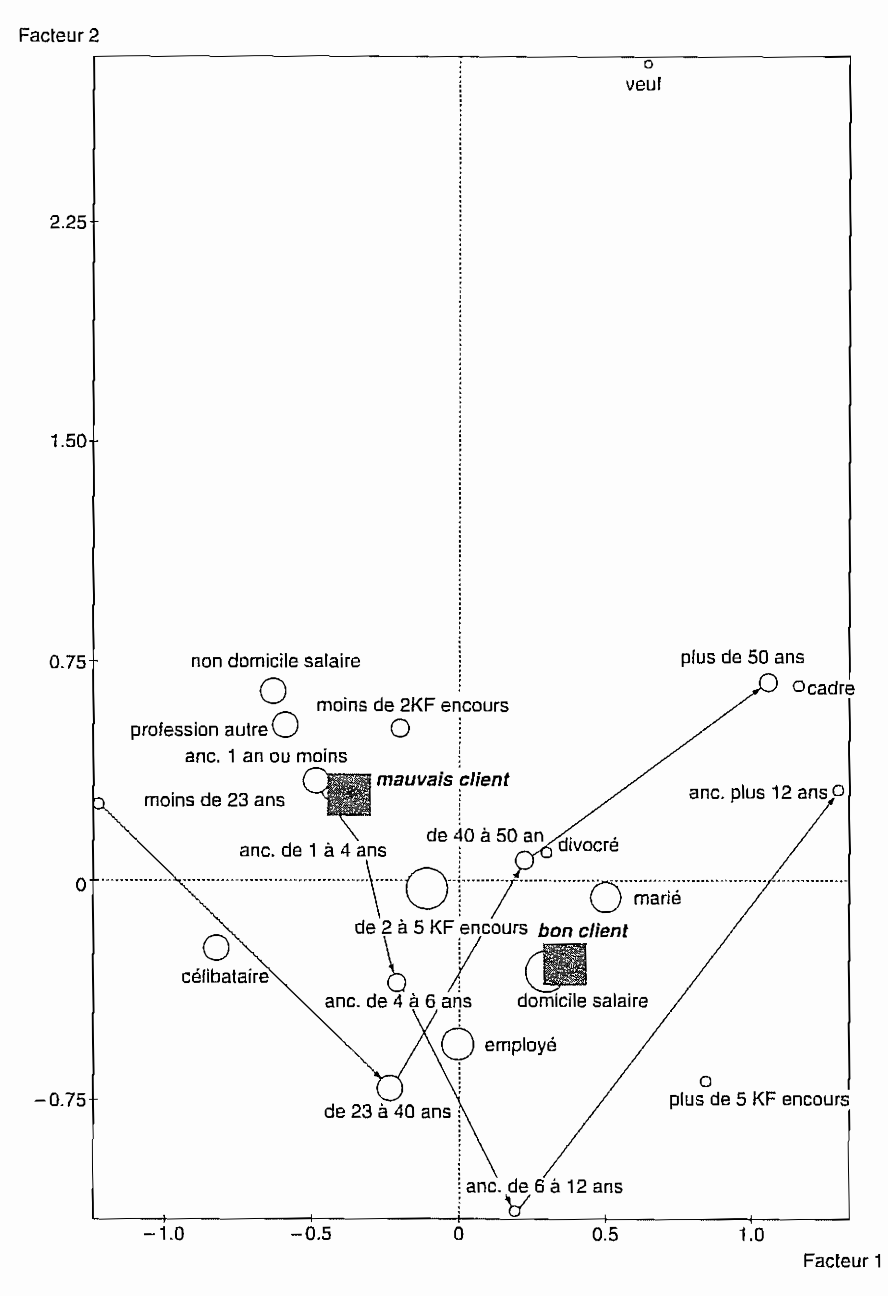

Facteur 2
2.25
1.50
0.75
0
−0.75
−1.0
−0.5
0
0.5
1.0
Facteur 1
veuf
non domicile salaire
profession autre
moins de 2KF encours
anc. 1 an ou moins
mauvais client
moins de 23 ans
anc. de 1 à 4 ans
de 40 à 50 an
divocré
plus de 50 ans
cadre
anc. plus 12 ans
marié
de 2 à 5 KF encours
bon client
célibataire
anc. de 4 à 6 ans
domicile salaire
employé
de 23 à 40 ans
plus de 5 KF encours
anc. de 6 à 12 ans

FIGURE 18.16

Tableau 18.7

MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
2 . Age du client													
AGE1 - moins de 23 ans	88	88.00	-12.8	2.8	4.8	9.6	-1.1	-1.23	0.27	0.46	0.92	-0.10	4.32
AGE2 - de 23 à 40 ans	150	150.00	-3.4	-10.5	3.6	-9.8	1.5	-0.23	-0.71	0.25	-0.66	0.10	2.12
AGE3 - de 40 à 50 ans	122	122.00	2.9	0.9	-15.5	-1.2	0.1	0.22	0.07	-1.21	-0.09	0.01	2.84
AGE4 - plus de 50 ans	108	108.00	12.6	8.1	7.7	3.2	-0.8	1.07	0.68	0.65	0.27	-0.07	3.33
3 . Situation familiale													
CELB - célibataire	170	170.00	-13.4	-3.7	7.2	6.7	-1.1	-0.82	-0.23	0.44	0.41	-0.07	1.75
MARI - marié	221	221.00	10.3	-1.1	-9.0	-0.2	8.6	0.50	-0.05	-0.44	-0.01	0.42	1.12
DIVO - divorcé	61	61.00	2.5	0.8	-1.6	-9.4	-12.1	0.30	0.10	-0.19	-1.12	-1.45	6.67
VEUF - veuf	16	16.00	2.6	11.3	8.7	0.3	1.7	0.65	2.79	2.14	0.07	0.42	28.25
4 . Ancienneté													
ANC1 - anc. 1 an ou moins	199	199.00	-9.0	6.3	-8.4	-0.8	-7.0	-0.49	0.34	-0.45	-0.04	-0.38	1.35
ANC2 - anc. de 1 à 4 ans	47	47.00	-2.3	0.7	2.2	-8.4	7.9	-0.32	0.09	0.31	-1.16	1.10	8.96
ANC3 - anc. de 4 à 6 ans	69	69.00	-1.9	-3.1	4.6	0.3	5.5	-0.21	-0.34	0.52	0.03	0.61	5.78
ANC4 - anc. de 6 à 12 ans	66	66.00	1.7	-10.0	6.9	1.0	-8.3	0.19	-1.14	0.79	0.11	-0.95	6.09
ANC5 - anc. plus 12 ans	87	87.00	13.4	3.2	-1.5	6.3	5.2	1.30	0.31	-0.14	0.61	0.50	4.38
5 . Domiciliation du salaire													
Soui - domicile salaire	316	316.00	9.4	-9.7	-0.8	6.4	2.5	0.30	-0.31	-0.03	0.21	0.08	0.48
Snon - non domicile salaire	152	152.00	-9.4	9.7	0.8	-6.4	-2.5	-0.63	0.64	0.05	-0.43	-0.17	2.08
7 . Profession													
CADR - cadre	77	77.00	11.2	6.4	5.2	-1.0	-6.8	1.17	0.66	0.54	-0.10	-0.71	5.08
EMPL - employé	237	237.00	0.0	-12.2	2.8	-1.6	6.5	0.00	-0.56	0.13	-0.07	0.29	0.97
AUTR - profession autre	154	154.00	-8.8	8.0	-7.0	2.4	-1.5	-0.58	0.53	-0.46	0.16	-0.10	2.04
8 . Moyenne en cours													
ENC1 - moins de 2KF encours	98	98.00	-2.2	5.8	4.4	-11.4	6.7	-0.20	0.52	0.40	-1.03	0.60	3.78
ENC2 - de 2 à 5 KF encours	308	308.00	-3.2	-0.8	-4.1	11.0	-0.4	-0.11	-0.03	-0.14	0.37	-0.01	0.52
ENC3 - plus de 5 KF encours	62	62.00	7.1	-5.8	0.5	-1.7	-7.5	0.85	-0.68	0.06	-0.20	-0.89	6.55
1 . Type de client													
BON - bon client	237	237.00	7.9	-6.2	-0.1	6.0	0.0	0.36	-0.28	-0.01	0.27	0.00	0.97
MAUV - mauvais client	231	231.00	-7.9	6.2	0.1	-6.0	0.0	-0.37	0.29	0.01	-0.28	0.00	1.03

TABLEAU 18.8

Facteurs	Corrélations avec la F.L.D. seuil = 0.093)	Coefficients de la F.L.D.	Probabilité
F 1	0.368	1.886240	0.0000
F 2	−0.289	−1.736910	0.0000
F 3	−0.005	−0.034836	0.8893
F 4	0.277	1.873010	0.0000
F 5	0.000	0.001670	0.9950
F 6	−0.011	−0.079781	0.7711
F 7	−0.060	−0.437118	0.1174
F 8	−0.094	−0.702389	0.0149
F 9	0.057	0.441749	0.1378
F 10	0.072	0.579235	0.0622
F 11	0.046	0.383689	0.2323
F 12	−0.096	−0.853755	0.0126
F 13	−0.009	−0.084949	0.8070
F 14	−0.077	−0.720454	0.0456
F 15	0.054	0.584582	0.1573
R2 = 0.33515	F = 15.19020		PROBA = 0.000
D2 = 2.00811	T2 = 234.91037		PROBA = 0.000

TABLEAU 18.9

Fonction linéaire de Fisher reconstituée à partir des variables d'origine

	Coefficients de la F.L.D.	Ecart-type bootstrap
Age du client		
moins de 23 ans	−1.311660	0.904747
de 23 à 40 ans	−0.461863	0.920698
de 40 à 50 ans	0.673484	0.932203
plus de 50 ans	0.949445	0.910071
Situation familiale		
célibataire	1.141380	0.697428
marié	0.341793	0.525616
divorcé	−2.254970	1.057110
veuf	−8.251150	2.230390
Ancienneté		
anc. 1 an ou moins	−4.034720	0.490477
anc. de 1 à 4 ans	−0.803805	1.830950
anc. de 4 à 6 ans	1.931500	0.630443
anc. de 6 à 12 ans	2.714630	1.091830
anc. plus 12 ans	6.071820	1.190080
Domiciliation du salaire		
domicile salaire	3.663660	0.538523
non domicile salaire	−7.616560	1.119560
Profession		
Cadre	3.846700	1.095720
Employé	0.062360	0.585443
profession autre	−2.019320	0.660032
Moyenne en cours		
moins de 2KF encours	−8.395870	1.134600
de 2 à 5 KF encours	1.929690	0.289948
plus de 5 KF encours	3.684670	0.939866

Le tableau 18.8 indique que les 15 dimensions ne sont pas toutes utiles. On élimine les facteurs n°3, 5, 6, 7, 9, 11, 13, 15, ce qui ramène à une discrimination dans un espace à 7 dimensions. Les composantes étant orthogonales, les coefficients ne changent pas après élimination (à un facteur près).

Le tableau 18.9 donne la grille de score brute avec une estimation par un bootstrap avec 1000 tirages des écart-types des coefficients ; rappelons qu'il n'y a en effet pas de formule permettant d'obtenir ces erreurs standard.

Le score d'un célibataire de 30 ans ayant un compte depuis 5 ans etc. s'obtient alors en effectuant la somme $-0.461863 + 1.141380 + 1.9315 + \ldots$

De telles valeurs ne sont pas commodes à utiliser et une pratique courante consiste à transformer linéairement les notes de score pour qu'elles soient comprises entre 0 et 1000. On ajoute aux coefficients de chaque variable une constante telle que la plus mauvaise note soit 0 : ici on ajoute +1.31166 aux modalités de « age du client » , +8.25115 aux modalités de « situation familiale » etc. On effectue ensuite une multiplication par une constante pour que le maximum soit de 1000. Le tableau 18.10 fournit ces scores après avoir réordonné variables et modalités selon l'amplitude de variation des scores pour mettre en évidence les variables et les modalités influentes.

TABLEAU 18.10

COEFFICIENTS REORDONNES DE LA FONCTION SCORE

IDEN LIBELLES	COEFFICIENTS DU SCORE	HISTOGRAMMES DES POINTS DE SCORE
8 . Moyenne en cours		
ENC3 - plus de 5 KF encours	236.93	********************
ENC2 - de 2 à 5 KF encours	202.51	*****************
ENC1 - moins de 2KF encours	0.00	*
5 . Domiciliation du salaire		
Soui - domicile salaire	221.24	*******************
Snon - non domicile salaire	0.00	*
4 . Ancienneté		
ANC5 - anc. plus 12 ans	198.22	*****************
ANC4 - anc. de 6 à 12 ans	132.37	***********
ANC3 - anc. de 4 à 6 ans	117.01	**********
ANC2 - anc. de 1 à 4 ans	63.37	*****
ANC1 - anc. 1 an ou moins	0.00	*
3 . Situation familiale		
CELB - célibataire	184.21	****************
MARI - marié	168.53	**************
DIVO - divorcé	117.60	**********
VEUF - veuf	0.00	*
7 . Profession		
CADR - cadre	115.05	**********
EMPL - employé	40.83	***
AUTR - profession autre	0.00	*
2 . Age du client		
AGE4 - plus de 50 ans	44.35	****
AGE3 - de 40 à 50 ans	38.93	***
AGE2 - de 23 à 40 ans	16.67	*
AGE1 - moins de 23 ans	0.00	*

En représentant simultanément les fonctions de répartition du score des deux groupes, il est alors possible de choisir des seuils de décision en fonction des risques de mauvaise classification, avec éventuellement une zone d'incertitude. La figure 18.17 illustre cette pratique : si l'on décide qu'un client ayant un score inférieur à 550 est « mauvais » on détecte environ 60 % de cette catégorie, tout en ne déclarant « mauvais » que 10 % des « bons ». Inversement si le seuil pour être classé « bon » est 750, on reconnaît environ la moitié de cette catégorie, et seuls 9.5 % des « mauvais » sont considérés à tort comme des « bons ».

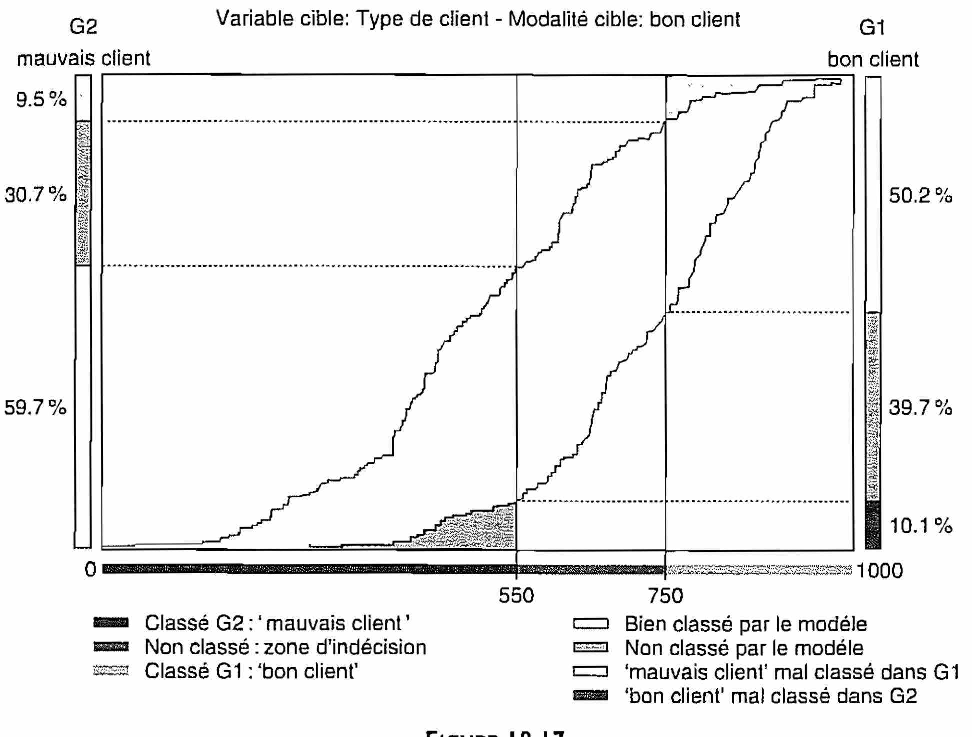


FIGURE 18.17

18.5 ANALYSE DISCRIMINANTE PROBABILISTE

18.5.1 La règle bayésienne et le modèle gaussien

Le modèle suivant fournit le cadre inférentiel nécessaire à l'analyse discriminante.

On suppose que les k groupes sont en proportion p_1 p_2, ..., p_k dans la population totale et que la distribution de probabilité du vecteur observation $\mathbf{x} = (x_1, ..., x_p)$ est donnée pour chaque groupe j par une densité (ou une loi discrète) $f_j(\mathbf{x})$.

Observant un point de coordonnées $(x_1, x_2, \ldots, x_p)$ la probabilité qu'il provienne du groupe j est donnée par la formule de Bayes :

$$P(G_j/\mathbf{x}) = \frac{p_j f_j(\mathbf{x})}{\sum_{j=1}^{k} p_j f_j(\mathbf{x})}$$

La règle bayésienne consiste alors à affecter l'observation $\mathbf{x}$ au groupe qui a la probabilité *a posteriori* maximale.

18.5.1.1 Le cas d'égalité des matrices de variance covariance

Si $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$, la règle devient linéaire. En effet ln (det Σ_j) est une constante et $(\mathbf{x} - \boldsymbol{\mu}_j)\,\Sigma^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_j)$ est alors égale à $\Delta^2(\mathbf{x}, \boldsymbol{\mu}_j)$, distance de Mahalanobis théorique de $\mathbf{x}$ à $\boldsymbol{\mu}_j$.

En développant et en éliminant $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ qui ne dépend pas du groupe on a :

$$\boxed{\max \left\{ \mathbf{x}'\Sigma^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j'\Sigma^{-1}\boldsymbol{\mu}_j + \ln p_j \right\}}$$

Si Σ est estimé par $\dfrac{n}{n-k}\mathbf{W}$, la règle bayésienne correspond à la règle géométrique lorsqu'il y a égalité des probabilités *a priori*. ***La règle géométrique est alors optimale.***

La probabilité *a posteriori* d'appartenance au groupe j est proportionnelle à :

$$p_j \exp\left(-\frac{1}{2}\Delta^2(\mathbf{x}, \boldsymbol{\mu}_j)\right)$$

Les dénominateurs étant les mêmes pour les k groupes on doit donc chercher le maximum de :

$$p_j f_j(\mathbf{x})$$

Il est donc nécessaire de connaître ou d'estimer $f_j(\mathbf{x})$. Diverses possibilités existent ; la plus classique étant de supposer que $\mathbf{x}$ suit une loi $N_p(\boldsymbol{\mu}, \Sigma_j)$ pour chaque groupe :

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}(\det \Sigma_j)^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)'\Sigma_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\right]$$

La règle bayésienne max $p_j f_j(\mathbf{x})$ revient donc en passant en logarithmes à minimiser :

$$(\mathbf{x} - \boldsymbol{\mu}_j)'\Sigma_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) - 2\ln p_j + \ln(\det \Sigma_j)$$

Lorsque les Σ_j sont différents cette règle est donc ***quadratique*** et il faut comparer k fonctions quadratiques de $\mathbf{x}$.

Σ_j est en général estimé par $\dfrac{n}{n-1}\mathbf{V}_j$ et $\boldsymbol{\mu}_j$ par $\mathbf{g}_j$.

18.5.1.2 Deux groupes avec égalité des matrices de variance

On affectera $\mathbf{x}$ au groupe 1 si :

$$\boxed{\mathbf{x}'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln\frac{p_2}{p_1}}$$

Si $p_1 = p_2 = 0.5$ on trouve la règle de Fisher en estimant Σ par $\dfrac{n}{n-2}\mathbf{W}$.

Soit :

$$\boxed{S(\mathbf{x}) = \mathbf{x}'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \ln\frac{p_2}{p_1}}$$

On affectera $\mathbf{x}$ au groupe 1 si $S(\mathbf{x}) > 0$ et au groupe 2 si $S(\mathbf{x}) < 0$.

La fonction $S(\mathbf{x})$ appelée *score* ou statistique d'Anderson est liée simplement à la probabilité *a posteriori* d'appartenance au groupe 1.

On a en effet :

$$P(G_1/\mathbf{x}) = P = \frac{p_1 f_1(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})}$$

d'où :

$$\frac{1}{P} = 1 + \frac{p_2 f_2(\mathbf{x})}{p_1 f_1(\mathbf{x})} = 1 + \frac{p_2}{p_1}\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right]$$

$$\frac{1}{P} - 1 = \frac{p_2}{p_1}\exp\left[\frac{1}{2}\Delta^2(\mathbf{x}; \boldsymbol{\mu}_1) - \frac{1}{2}\Delta^2(\mathbf{x}; \boldsymbol{\mu}_2)\right]$$

d'où $\ln\left(\dfrac{1}{P} - 1\right) = -S(\mathbf{x})$.

Soit :

$$\boxed{P(G/\mathbf{x}) = \frac{1}{1 + \exp(-S(\mathbf{x}))} = \frac{\exp(S(\mathbf{x}))}{1 + \exp(S(\mathbf{x}))}}$$

P est donc fonction logistique du *score*.

Lorsque $p_1 = p_2 = 1/2$:

$$P = \frac{1}{1 + \exp\left(-\frac{1}{2}(\Delta^2(\mathbf{x}; \boldsymbol{\mu}_1) - \Delta^2(\mathbf{x}; \boldsymbol{\mu}_2))\right)}$$

on remarque que modifier les probabilités *a priori* se traduit simplement par un changement du terme constant. Dans de nombreuses aplications, ce qui compte essentiellement est la combinaison linéaire des variables, le terme constant étant laissé au choix du praticien (voir plus haut). Dans

ces conditions, le fait que les proportions des groupes soient conformes ou non à la réalité devient sans importance, ce qui compte pour la qualité des estimations étant la taille des échantillons.

Voici à titre d'exemple le tableau 18.11 donnant les affectations des 45 premières observations des données d'infarctus selon la règle précédente. L'astérisque indique une erreur de classement.

TABLEAU 18.11

	Groupe réel	Groupe attribué		P(G₁/x)	P(G₂/x)
1	SURVIE	SURVIE		0.4515	0.5485
2	DECES	DECES		0.8140	0.1860
3	DECES	DECES		0.9597	0.0403
4	SURVIE	SURVIE		0.2250	0.7750
5	DECES	DECES		0.8112	0.1888
6	DECES	DECES		0.8928	0.1072
7	SURVIE	SURVIE		0.3202	0.6798
8	SURVIE	DECES	*	0.8711	0.1289
9	SURVIE	SURVIE		0.0984	0.9016
10	SURVIE	SURVIE		0.0797	0.9203
11	SURVIE	SURVIE		0.0138	0.9862
12	SURVIE	SURVIE		0.0160	0.9840
13	SURVIE	SURVIE		0.0052	0.9948
14	SURVIE	SURVIE		0.0105	0.9895
15	SURVIE	SURVIE		0.0019	0.9981
16	SURVIE	SURVIE		0.0258	0.9742
17	SURVIE	SURVIE		0.2011	0.7989
18	SURVIE	SURVIE		0.2260	0.7740
19	SURVIE	SURVIE		0.0022	0.9978
20	SURVIE	SURVIE		0.1222	0.8778
21	SURVIE	SURVIE		0.0014	0.9986
22	DECES	DECES		0.8629	0.1371
23	DECES	SURVIE	*	0.4804	0.5196
24	DECES	DECES		0.9900	0.0100
25	DECES	DECES		0.5845	0.4155
26	DECES	DECES		0.7447	0.2553
27	DECES	DECES		0.7067	0.2933
28	DECES	SURVIE	*	0.4303	0.5697
29	SURVIE	SURVIE		0.1118	0.8882
30	SURVIE	DECES	*	0.5734	0.4266
31	SURVIE	SURVIE		0.2124	0.7876
32	DECES	DECES		0.9928	0.0072
33	DECES	DECES		0.7301	0.2699
34	SURVIE	DECES	*	0.5354	0.4646
35	DECES	DECES		0.9943	0.0057
36	SURVIE	SURVIE		0.1218	0.8782
37	SURVIE	SURVIE		0.2757	0.7243
38	SURVIE	SURVIE		0.1759	0.8241
39	DECES	DECES		0.9555	0.0445
40	SURVIE	SURVIE		0.0695	0.9305
41	DECES	DECES		0.9762	0.0238
42	DECES	DECES		0.9785	0.0215
43	SURVIE	SURVIE		0.3240	0.6760
44	SURVIE	SURVIE		0.2121	0.7879
45	DECES	DECES		0.7880	0.2120

Dans l'exemple infarctus, le logiciel a supposé par défaut l'égalité des probabilités *a priori*, ce qui est contestable. Les probabilités *a posteriori* sont donc dépendantes de cette hypothèse.

Sous réserve du caractère réaliste de l'hypothèse de multinormalité, ces résultats sont donc plus précis qu'une simple décision selon la distance la plus courte. Le calcul de probabilité *a posteriori* montre ici que 4 classements erronés sur 5 se sont produits dans une zone d'incertitude (probabilités voisines de 0.5).

18.5.1.3 Taux d'erreur théorique pour deux groupes avec $\Sigma_1 = \Sigma_2$

Quand $p_1 = p_2$, la règle de classement théorique est d'affecter au groupe 1 si :

$$S(\mathbf{x}) = \mathbf{x}'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$$

La probabilité d'erreur de classement est donc :

$$P(S(\mathbf{x}) > 0/\mathbf{x} \in N_p(\boldsymbol{\mu}_2; \Sigma))$$

La loi de $S(\mathbf{x})$ est une loi de Gauss à 1 dimension comme combinaison linéaire des composantes de $\mathbf{x}$.

$$E(S(\mathbf{x})) = \boldsymbol{\mu}'_2\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2}\Delta_p^2$$

$$V(S(\mathbf{x})) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Delta_p^2$$

d'où :

$$S(\mathbf{x}) \text{ suit une LG}\left(-\frac{1}{2}\Delta_p^2; \Delta_p\right) \quad \text{si} \quad \mathbf{x} \in G_2$$

La probabilité de classer dans le groupe 1 une observation du groupe 2 est :

$$\boxed{P(1/2) = P\left(U > \frac{\Delta_p}{2}\right)}$$

Elle est égale à $P(2/1)$. Cette relation donne une interprétation concrète à la distance de Mahalanobis.

Si $p_1 \neq p_2$ on trouve :

$$P(1/2) = P\left(U > \frac{\Delta_p}{2} + \frac{1}{\Delta_p}\ln\frac{p_2}{p_1}\right)$$

$$P(2/1) = P\left(U > \frac{\Delta_p}{2} - \frac{1}{\Delta_p}\ln\frac{p_2}{p_1}\right)$$

Lorsque $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$ sont estimés, $S(\mathbf{x})$ ne suit plus une loi normale et utiliser D_p comme estimation de Δ_p conduit à une estimation biaisée des probabilités d'erreur de classement : il

y a en moyenne sous-estimation de la probabilité globale d'erreur $p_1 P(2/1) + p_2P(1/2)$, due entre autres raisons au fait que D_p^2 surestime Δ_p^2 (voir chapitre 15, paragraphe 15.5.6C).

Pour l'exemple des infarctus comme $D_p = 2.223$ on aboutit à une estimation du taux d'erreur égale à $P(U > 1.11) = 0.13$.

L'utilisation de l'estimation sans biais de Δ^2, $\frac{n-p-1}{n-2}D^2 - p\frac{n}{n_1 n_2} = 4.37$ conduit à une estimation du taux d'erreur voisine de 15 %.

La règle bayésienne peut cependant conduire à des décisions absurdes lorsque les probabilités *a priori* sont très déséquilibrées. Supposant par exemple que $p_1 = 0.01$ et $p_2 = 0.99$, ce qui correspond à la détection d'un groupe rare. Il est alors facile de voir que pratiquement toutes les observations seront classées en *G2* et aucune en *G1*. Notons qu'une telle règle donne un pourcentage global de bons classement de 99 % ! (moyenne de 0 % sur le groupe 1 et 100 % sur le groupe 2).

En effet, si l'on écrit la formule de Bayes sous la forme $P(G1/\mathbf{x}) = p_1 \frac{f_1(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})}$, il faudrait que $\frac{f_1(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})} > 50$ pour que la probabilité *a posteriori* dépasse 0.5, ce qui n'est possible que si $f_1(\mathbf{x}) > 99 f_2(\mathbf{x})$, ce qui est fort improbable.

18.5.1.4 Tests et sélection de variables

L'hypothèse d'égalité des matrices $\boldsymbol{\Sigma}_i$ peut être testée au moyen du test de Box qui généralise celui de Bartlett pour le cas unidimensionnel.

Si l'hypothèse $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_k$ est vraie, la quantité :

$$\left(1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right)\left[\left(\sum_i \frac{1}{n_i - 1} - \frac{1}{n-k}\right)(n-k)\ln\left|\frac{n}{n-k}\mathbf{W}\right| - \sum_i (n_i - 1)\ln\left|\frac{n_i}{n_i - 1}\mathbf{V}_i\right|\right]$$

suit approximativement une loi χ^2 *à* $\frac{p(p+1)(k-1)}{2}$ degrés de liberté.

Si l'on rejette l'hypothèse d'égalité, doit-on pour autant utiliser les règles quadratiques ? Cela n'est pas sûr dans tous les cas. Tout d'abord le test de Box n'est pas parfaitement fiable, ensuite l'usage de règles quadratiques implique l'estimation de bien plus de paramètres que la règle linéaire, puisqu'il faut estimer chaque $\boldsymbol{\Sigma}_j$. Lorsque les échantillons sont de petite taille, les fonctions obtenues sont très peu robustes et il vaut mieux utiliser une règle linéaire malgré tout.

Pour deux groupes le résultat suivant est à l'origine des méthodes classiques de sélection de variables :

Soit un sous-ensemble de l variables parmi les p composantes de $\mathbf{x}$.

Supposons que $\Delta_p^2 = \Delta_l^2$; en d'autres termes les $p - l$ variables restantes n'apportent aucune information pour séparer les deux populations; alors :

$$\frac{(n_1 + n_2 - p - 1)n_1 n_2(D_p^2 - D_l^2)}{(p - l)(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_l^2} = F(p - l; n_1 + n_2 - p - 1)$$

On peut ainsi tester l'accroissement de la distance de Mahalanobis apporté par une nouvelle variable à un groupe déjà constitué en prenant $l = p - 1$.

Lorsque l'on fait de la discrimination entre plus de deux groupes, les tests sont ceux utilisant le Λ de Wilks.

Le test d'égalité des k espérances $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$ est le suivant :

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{V}|} = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|} = \frac{1}{|\mathbf{W}^{-1}\mathbf{B} + \mathbf{I}|}$$

suit la loi de Wilks de paramètres $p, n - k, k - 1$ sous $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$

car $n\mathbf{V}$, $n\mathbf{W}$, $n\mathbf{B}$ suivent respectivement les lois de Wishart à $n - 1$, $n - k$, $k - 1$ degrés de liberté.

Si $k = 3$ on utilisera la loi exacte de Λ et non une approximation :

$$\frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} = \frac{p}{n - p - 2} F(2p; 2(n - p - 2)$$

Si $k = 2$, le test de Wilks et le test de la distance de Mahalanobis ($H_0 : \Delta_p^2 = 0$) sont identiques car $\mathbf{B}$ étant de rang 1, on a :

$$\Lambda = \frac{1}{1 + D_p^2 \dfrac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2)}} = \frac{1}{\mu + 1} = 1 - \lambda$$

Le test de $H_0 : \boldsymbol{\mu}_i = \boldsymbol{\mu} \forall i$ peut s'effectuer également en utilisant comme statistique de test la trace de $\mathbf{W}^{-1}\mathbf{B}$ appelée statistique de Lawley-Hotelling qui suit la loi du T_0^2 généralisé de Hotelling approximable par un $\chi^2_{p(k-1)}$.

La trace de $\mathbf{V}^{-1}\mathbf{B}$ est appelée trace de Pillai. Pour l'introduction pas à pas de variables en discriminante à k groupes on utilise souvent le test de variation de Λ mesuré par :

$$\frac{n - k - p}{k - 1}\left(\frac{\Lambda_p}{\Lambda_{p+1}} - 1\right)$$

que l'on compare à un $F_{k-1; n-k-p}$.

Comme en régression multiple, il existe divers algorithmes de sélection : ascendant, descendant etc. D'ailleurs pour deux groupes, les méthodes sont identiques (voir paragraphe 18.2.3).

L'application d'une méthode ascendante aux données « infarctus » conduit aux résultats suivants :

The STEPDISC Procedure
Forward Selection: Step 1

Statistics for Entry, DF = 1, 99

Variable	R-Square	F Value	Pr > F	Tolerance
FRCAR	0.0535	5.60	0.0200	1.0000
INCAR	0.4826	92.33	<.0001	1.0000
INSYS	0.4493	80.75	<.0001	1.0000
PRDIA	0.2228	28.37	<.0001	1.0000
PAPUL	0.1844	22.38	<.0001	1.0000
PVENT	0.0719	7.67	0.0067	1.0000
REPUL	0.4198	71.62	<.0001	1.0000

La variable INCAR est alors sélectionnée car la la plus explicative.

Forward Selection: Step 2

Statistics for Entry, DF = 1, 98

Variable	Partial R-Square	F Value	Pr > F	Tolerance
FRCAR	0.0461	4.73	0.0320	0.9874
INSYS	0.0265	2.66	0.1058	0.2130
PRDIA	0.1089	11.98	0.0008	0.8699
PAPUL	0.1223	13.66	0.0004	0.9274
PVENT	0.0110	1.09	0.2994	0.9205
REPUL	0.0622	6.50	0.0124	0.4117

C'est ensuite la variable PAPUL et la sélection s'arrête car plus aucune variable n'est significative au pas n°3 conditionnellement aux choix précédents.

Forward Selection: Step 3

Statistics for Entry, DF = 1, 97

Variable	Partial R-Square	F Value	Pr > F	Tolerance
FRCAR	0.0107	1.05	0.3090	0.8104
INSYS	0.0013	0.13	0.7197	0.1832
PRDIA	0.0003	0.03	0.8545	0.1259
PVENT	0.0020	0.19	0.6609	0.8777
REPUsL	0.0000	0.00	0.9784	0.1994

18.5.2 Méthodes « non paramétriques »

On ne fait pas d'hypothèse spécifique sur la famille de loi de probabilité.

Des variantes multidimensionnelles de la méthode du noyau permettent d'estimer $f_j(\mathbf{x})$.

$$\hat{f}_j(\mathbf{x}) = \frac{1}{n_j h} \sum_{i=1}^{n_j} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

où K est une densité multidimensionnelle.

La discrimination « par boules » en est un cas particulier : on trace autour de $\mathbf{x}$ une boule de rayon ρ donné dans $\mathbb{R}^p$ et on compte le nombre d'observation k_j du groupe j dans cette boule. On estimera alors directement $P(G_j/\mathbf{x})$ par :

$$\frac{k_j}{\sum_j k_j}$$

Remarque : La boule peut être vide si ρ est trop petit.

Une des méthodes les plus utilisées est cependant la méthode des k plus proches voisins. On cherche les k points les plus proches de $\mathbf{x}$ au sens d'une métrique à préciser et on classe $\mathbf{x}$ dans le groupe le plus représenté : la probabilité *a posteriori* s'obtient comme pour la discrimination par boules mais n'a pas grand sens si k est faible.

La méthode du noyau est en théorie optimale, mais est cependant peu utilisée car le réglage des paramètres de lissage est assez délicat.

18.6 RÉGRESSION LOGISTIQUE BINAIRE (DEUX GROUPES)

Au paragraphe 18.5.1.2 on a établi sous les hypothèses de normalité et égalité des matrices de variance covariance que la probabilité *a posteriori* d'appartenance au groupe 1 se mettait sous la forme d'une fonction logistique du score, lui-même combinaison linéaire des variables. La régression logistique, appelée également modèle "logit", consiste à poser cette relation comme hypothèse de départ, ce qui est donc un modèle plus large que celui de l'analyse discriminante probabiliste. La régression logistique a été introduite en 1944 par Berkson en biostatistique, puis en 1973 par McFadden en économètrie.

$$P(G1/\mathbf{x}) = \pi(\mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} = \frac{e^{\beta_0 + \boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\beta_0 + \boldsymbol{\beta}'\mathbf{x}}}$$

Ce modèle est souvent qualifié de semi-paramétrique, dans la mesure où on modélise le rapport des densités $\dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})}$ au lieu de chacune. On notera $P(G1/\mathbf{x}) = P(Y = 1)$.

Pour des compléments, en particulier pour le cas polytomique, on se reportera à J.J. Droesbeke et al. (2005).

18.6.1 Interprétation

Le choix de la fonction logistique conduit à une expression comprise entre 0 et 1, ce qui convient à une probabilité, et correspond souvent à une bonne représentation de certains phénomènes.

Les coefficients du modèle sont liés aux odds-ratios ou « rapport de cotes » de la manière suivante.

Considérons tout d'abord le cas d'une seule variable explicative binaire. Par exemple $x = 1$ si l'on fume, $x = 0$ sinon et $Y = 1$ désigne la survenance d'une maladie.

La probabilité d'être malade si l'on fume est $P(Y = 1/x = 1) = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}$ que l'on compare tout d'abord à la probabilité de ne pas être malade si l'on fume :

$$P(Y = 0/x = 1) = 1 - P(Y = 1/x = 1) = \frac{1}{1 + e^{\beta_0+\beta_1}}$$

L'odds est le rapport de ces deux probabilités $P(Y = 1/x = 1)/P(Y = 0/x = 1)$ analogue à la « cote » des parieurs.

On effectue ensuite les mêmes calculs pour les non fumeurs : la probabilité d'être malade est $P(Y = 1/X = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$, celle de ne pas être malade $P(Y = 1/X = 0) = \frac{1}{1 + e^{\beta_0}}$

L'odds ratio est alors : $OR = \frac{P(Y = 1/x = 1)/P(Y = 0/x = 1)}{P(Y = 1/x = 0)/P(Y = 0/x = 0)} = e^{\beta_1}$, c'est le facteur par lequel la cote est multipliée lorsque x passe de 0 à 1. l'OR est supérieur à 1 s'il y a aggravation.

Plus généralement pour une variable explicative numérique, on a :

$$OR = \frac{\pi(x + 1)/(1 - \pi(x + 1))}{\pi(x)/(1 - \pi(x))} = e^{\beta_1}$$

mais l'interprétation et la valeur de l'odds ratio dépendent de l'unité de mesure de la variable : si x désigne la quantité quotidienne de tabac, le rapport ne sera pas le même selon que x s'exprime en nombre de cigarettes, ou en nombre de paquets. Comme en régression linéaire, le produit βx reste fixe.

On peut sans difficulté utiliser des prédicteurs qualitatifs de la même manière que dans le modèle linéaire général. Chaque variable qualitative à m modalités est remplacée par $m - 1$ indicatrices après élimination d'une des modalités, dite modalité de référence, qui aura un coefficient nul. Les comparaisons de coefficients se font alors par rapport à cette modalité : une valeur proche de zéro ne signifie pas qu'une modalité est sans effet, mais qu'elle est proche de la modalité de référence.

18.6.2 Estimation

Elle s'effectue par la méthode du maximum de vraisemblance à partir d'un échantillon *iid* de n observations $(y_i, \mathbf{x}_i)$ prélevées dans la population totale. La vraisemblance correspond d'habitude à la probabilité d'observer les $(y_i, \mathbf{x}_i)$ mais il s'agit ici d'une vraisemblance conditionnelle puisque l'on ne modélise que $\pi(\mathbf{x})$:

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\beta_0+\boldsymbol{\beta}'\mathbf{x}}}\right)^{y_i} \left(1 - \frac{e^{\beta_0+\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\beta_0+\boldsymbol{\beta}'\mathbf{x}}}\right)^{1-y_i} = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i}$$

Notons $\pi_i = \pi(\mathbf{x}_i)$. En annulant les dérivées par rapport aux β_j de la log-vraisemblance :

$$\ell(\beta_0, \boldsymbol{\beta}) = \log L(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i)\log(1 - \pi_i)]$$

On aboutit au système d'équations :

$$\begin{cases} \dfrac{\partial \ell(\beta)}{\partial \beta_0} = \displaystyle\sum_{i=1}^{n}(y_i - \pi_i) = 0 \\ \dfrac{\partial \ell(\beta)}{\partial \beta_j} = \displaystyle\sum_{i=1}^{n} x_i^j(y_i - \pi_i) = 0 \qquad j = 1, \ldots, p \end{cases}$$

qui n'a pas de solution analytique et se résout par des procédures de calcul numérique.

On obtient la matrice de variance-covariance asymptotique des estimateurs, d'où les erreurs standard des coefficients, en appliquant les résultats du chapitre 13, paragraphe 13.4 par inversion de la matrice d'information de Fisher :

$$\hat{V}(\hat{\beta}) = \left[\frac{-\partial^2 \ell(\beta)}{\partial \beta^2}\right]^{-1}_{\beta=\hat{\beta}} = \begin{bmatrix} \displaystyle\sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i) & & \displaystyle\sum_{i=1}^{n} x_i^p \hat{\pi}_i(1 - \hat{\pi}_i) \\ & \ddots & \\ \displaystyle\sum_{i=1}^{n} x_i^p \hat{\pi}_i(1 - \hat{\pi}_i) & & \displaystyle\sum_{i=1}^{n} (x_i^p)^2 \hat{\pi}_i(1 - \hat{\pi}_i) \end{bmatrix}^{-1}$$

$$= \left(\begin{bmatrix} 1.. & x_1^p \\ \vdots & \vdots \\ 1.. & x_n^p \end{bmatrix}' \begin{bmatrix} \hat{\pi}_1(1 - \hat{\pi}_1) & & 0 \\ & \ddots & \\ 0 & & \hat{\pi}_n(1 - \hat{\pi}_n) \end{bmatrix} \begin{bmatrix} 1.. & x_1^p \\ \vdots & \vdots \\ 1.. & x_n^p \end{bmatrix}\right)^{-1}$$

$$= (\mathbf{X}'\hat{V}\mathbf{X})^{-1}$$

Le tableau 18.11 donne les résultats de la procédure Logistic de SAS pour les données infarctus (on modélise la probabilté de décès). Le khi-2 de Wald est égal au carré du rapport du coefficient estimé à son erreur standard estimée : il est analogue au carré du T de Student de la régression linéaire multiple.

Aucun coefficient n'apparaît significatif, ce qui s'explique par un phénomène de multi-colinéarité marqué.

Les estimations précédentes supposent un échantillonnage aléatoire simple dans une population avec pour conséquences que les effectifs observés de $G1$ et $G2$ sont aléatoires d'espérances respectives np_1 et np_2. Dans de nombreuses applications pratiques on utilise un échantillonnage stratifié (cf chapitre 20) où n_1 et n_2 sont fixés et où les proportions

TABLEAU 18.11

Paramètre	DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2	Odds ratio	Limites de confiance à 95 %	
Intercept	1	-1.3378	9.5500	0.0196	0.8886			
FRCAR	1	0.0474	0.0899	0.2786	0.5976	1.049	0.879	1.251
INCAR	1	-5.7825	5.3189	1.1819	0.2770	0.003	<0.001	103.790
INSYS	1	0.1102	0.3935	0.0784	0.7795	1.116	0.516	2.414
PRDIA	1	0.0390	0.1950	0.0401	0.8414	1.040	0.709	1.524
PAPUL	1	0.1511	0.2331	0.4199	0.5170	1.163	0.736	1.837
PVENT	1	0.0542	0.0789	0.4719	0.4921	1.056	0.904	1.232
REPUL	1	-0.0001	0.0039	0.0000	0.9978	1.000	0.992	1.008

p_1 et p_2 ne sont pas respectées : en particulier il est courant de sur-représenter le groupe le plus rare. On montre alors, comme pour l'analyse discriminante, que seule change l'estimation du terme constant β_0 à laquelle il suffit d'ajouter $\ln\left(\frac{p_1}{p_2}\right)$. On prendra garde que si les probabilités *a priori* sont inconnues le terme constant ne pourra être estimé et que donc les probabilités *a posteriori* seront incorrectes (définies à une transformation monotone près). S'il agit seulement de calculer un score de risque, c'est sans gravité.

18.6.3 Tests et sélection de variables

Trois méthodes sont disponibles pour tester l'apport d'une variable au modèle :

- Le test de Wald, déjà présenté,
- Le test du rapport des vraisemblances qui consiste à calculer pour chaque variable

$$-2\ln\left(\frac{\text{Vraisemblance sans la variable}}{\text{Vraisemblance avec la variable}}\right)$$

- Le test du score $U(\beta)'_{\hat{\beta}_{H_0}}[J(\hat{\beta}_{H_0})]^{-1}\,U(\beta)_{\hat{\beta}_{H_0}}$ où J est la matrice d'information de Fisher et U le vecteur des dérivées partielles de la log-vraisemblance estimés sous la contrainte $\beta_j = 0$. En régression logistique simple, le score est égal à nr^2, où r est le coefficient de corrélation linéaire (abusif!) entre Y et x

Ces trois tests suivent asymptotiquement un khi-deux à un degré de liberté sous l'hypothèse de nullité du coefficient théorique. La figure 18.18 illustre le comportement de la log-vraisemblance et permet de comparer ces trois tests qui donnent en général des résultats équivalents : le test de Wald compare l'écart entre le coefficient théorique et sa valeur estimée en abscisse, le test du rapport des vraisemblances compare, la différence en ordonnée et le test du score compare à zéro la pente de la tangente au point théorique.

Ces tests peuvent être utilisés pour des algorithmes de sélection (ascendante, descendante ou complète). La liste des meilleurs modèles (selon la valeur du khi-deux associé au score) de une à 7 variables est donnée dans le tableau 18.12.

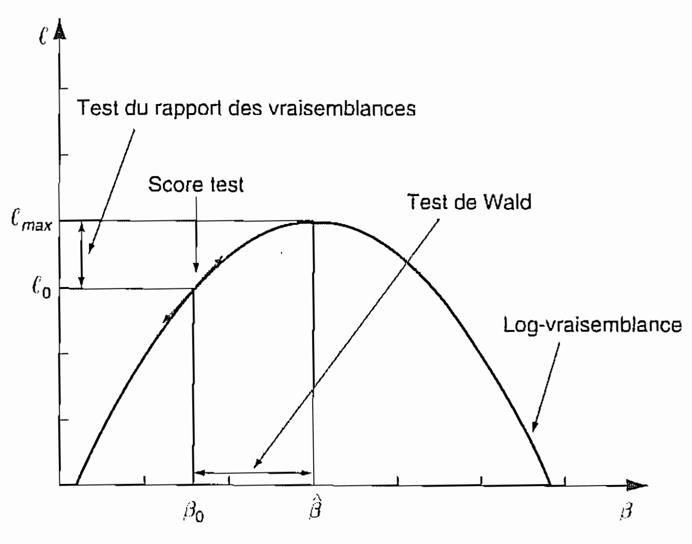


FIGURE 18.18

TABLEAU 18.12

Nombre de variables	Khi 2	Variables incluses dans le modèle
1	48.7385	INCAR
2	55.1304	INCAR PAPUL
3	55.6196	FRCAR INCAR PAPUL
4	56.1043	FRCAR INCAR INSYS PAPUL
5	56.2861	FRCAR INCAR INSYS PAPUL PVENT
6	56.3087	FRCAR INCAR INSYS PAPUL PVENT REPUL
7	56.3169	FRCAR INCAR INSYS PRDIA PAPUL PVENT REPUL

On retiendra le modèle à deux variables INCAR PAPUL, les deux mêmes qui avaient été sélectionnées en analyse discriminante d'où le modèle :

Paramètre	DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2	Odds ratio	Limites de confiance à 95%	
Intercept	1	2.9331	1.7855	2.6985	0.1004			
INCAR	1	−4.5491	0.9402	23.4083	<.0001	0.011	0.002	0.067
PAPUL	1	0.2015	0.0622	10.4937	0.0012	1.223	1.083	1.382

Ces tests peuvent servir à valider globalement un modèle c'est à dire à tester la nullité simultanée de tous les coefficients β (sauf de la constante). Ainsi la vraisemblance en l'absence

d'effet des p variables vaut $\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}$ puisque $P(Y=1)$ ne dépend plus des x et s'estime, dans le cas d'échantillonnage global *iid* par la proportion d'observations de $G1$.

On comparera $-2\ln L + 2\ln\left(\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}\right)$ à un khi-deux à p degrés de liberté.

18.6.4 Comparaison avec l'analyse discriminante linéaire

La régression logistique est souvent opposé à tort à l'analyse linéaire discriminante de Fisher, certains praticiens croyant que la régression logistique serait plus « scientifique » ; l'incompréhension de la propriété indiquée en 18.2.3 qui montre que la fonction de Fisher peut s'obtenir à l'aide d'une régression ordinaire a également beaucoup joué. En réalité si les deux modèles diffèrent par leurs hypothèses, ils sont en réalité assez proches et les scores obtenus sont dans la pratique très voisins. A titre d'exemple le coefficient de corrélation entre les scores de Fisher et logistique dans le cas infarctus vaut 0.99881 . . .

Les deux méthodes reposent sur des modèles probabilistes spécifiant des lois conditionnelles : les lois de $Y/\mathbf{x}$ pour la logistique, les lois de $\mathbf{x}/Y$ pour l'analyse discriminante.

Les deux modèles aboutissent à une fonction de score linéaire du même type $S(\mathbf{x}) = \beta_0 + \beta'\mathbf{x}$. La différence réside dans le mode d'estimation : maximum de vraisemblance pour la logistique, moindres carrés pour Fisher. Précisons d'ailleurs que si f_1 et f_2 sont des lois normales de même matrice de variance, alors l'estimation des moindres carrés donnée par la fonction de Fisher coïncide avec l'estimateur du maximum de vraisemblance complet et est donc plus précise que l'estimation fournie par la régression logistique qui n'est qu'un maximum de vraisemblance conditionnel qui ignore l'information sur les distributions des x.

En théorie la régression logistique est mieux adaptée au cas non-gaussien. Cependant la fonction de Fisher ayant aussi une justification géométrique peut être appliquée en dehors de tout contexte probabiliste.

Les coefficients sont déterminés de façon unique et ont une interprétation en termes d'odds ratio pour la logistique, alors que pour la fonction de Fisher, ils sont définis à un facteur multiplicatif près. Il existe cependant un cas où la régression logistique ne fonctionne pas, celui de la séparation linéaire complète : les estimateurs n'existent pas (non-convergence), défaut que ne possède la fonction de Fisher que dans le cas fort rare où les deux centres de gravité sont confondus.

Le fait que les erreurs-standard ne sont pas calculables en analyse discriminante alors qu'elles le sont en logistique est un argument en faveur de cette dernière, encore faut-il préciser qu'elles sont asymptotiques et que le bootstrap peut fournir des erreurs standard en discriminante.

La régression logistique a été conçue plus comme un modèle permettant de mettre en évidence des facteurs influents que comme une technique décisionnelle de prévision individuelle. Dans certaines applications (credit scoring en particulier), le score $\beta_0 + \beta'\mathbf{x}$ est utilisé à des fins de prévision : si l'objectif est purement opérationnel, il convient alors de choisir entre les méthodes en termes de pouvoir prédictif ou taux d'erreur, et non selon la valeur de statistiques de test, (voir plus loin), ni selon des présupposés idéologiques.

18.7 VALIDATION

La qualité d'un score ou d'une règle de classement n'est pas seulement un problème de test statistique, ou d'estimation d'une distance de Mahalanobis. En effet les statistiques de tests, pour utiles qu'elles soient, ne sont pas directement liées aux performances en termes de classement et reposent sur des hypothèses pas toujours vérifiées. Il faut non seulement définir des indicateurs pertinents, mais aussi pouvoir comparer différentes méthodes à l'aide de ces indicateurs. La comparaison de performances ne va pas toujours de soi, quand des modèles n'ont pas le même nombre de paramètres : le modèle le plus complexe sera plus performant sur les données qui ont servi à l'estimer, mais cela sera souvent trompeur. Il faut donc comparer les capacités prédictives sur de nouvelles données (ou observations supplémentaires), ce qui conduit à partager les données dont on dispose en plusieurs sous-échantillons. Le chapitre suivant reprendra ce problème sous un point de vue plus général.

18.7.1 Procédure de classement

Quelle que soit la méthode (discrimination linéaire ou quadratique, logistique, SVM etc.) si l'objectif est de prédire l'appartenance à des classes, les résultats finaux se présenteront sous forme d'un tableau de classement ou matrice de confusion obtenue en appliquant la méthode à des observations dont l'appartenance est connue et comparée à l'appartenance prédite (voir 18.1.3).

Ainsi pour les données infarctus, en utilisant la fonction de Fisher avec les 7 variables, et en utilisant la règle bayésienne avec égalité des probabilités *a priori*, on obtient 87 % d'observations bien classées :

De PRONO	DECES	SURVIE	Total
DECES	46	5	51
	90.20	9.80	100.00
SURVIE	8	42	50
	16.00	84.00	100.00
Total	54	47	101
	53.47	46.53	100.00

Or si l'on se contente de classer les observations qui ont permis d'estimer le modèle (« resubstitution ») on commet une erreur méthodologique qui peut-être grave si la taille des échantillons est peu élevée (jusqu'à quelques centaines) et le modèle complexe. En effet on aura tendance à trouver des résultats flatteurs puisque l'on utilise deux fois les mêmes données, une fois pour estimer les paramètres du modèle et leur donner donc les meilleures valeurs possibles, et encore une fois pour classer les données. Un modèle à 50 paramètres donnera toujours un excellent ajustement, mais se révelera inefficace à l'avenir. La capacité prédictive ne peut se juger que sur des données indépendantes.

On recommande donc de séparer aléatoirement les données en deux ensembles dits d'apprentissage et de test. L'ensemble d'apprentissage sert à estimer un modèle qui va être utilisé sur l'ensemble test.

Ceci n'est toutefois pas suffisant et pour obtenir non pas une seule estimation du taux de bien classés, mais également un intervalle de confiance, il faut répéter le tirage aléatoire plusieurs fois. On recommande d'effectuer un tirage stratifié dans chaque groupe pour éviter des fluctuations parasites des effectifs des groupes.

Lorsque le nombre d'observations disponibles est faible, comme dans le cas des infarctus, il n'est pas possible de séparer les données en deux sous-ensembles. On utilise alors la validation croisée qui consiste à effectuer n analyses discriminantes : on ôte tour à tour chaque observation que l'on prédit à l'aide d'un modèle estimé sur les $n-1$ observations restantes. Cette méthode s'apparente donc au jack-knife et au calcul du « *press* » en régression. On obtient des estimations de biais faible, voire nul, mais avec une variance pas toujours négligeable.

Voici le résultat pour les données infarctus : l'estimation du taux de bons classements diminue à 84 %

De PRONO	DECES	SURVIE	Total
DECES	44	7	51
	86.27	13.73	100.00
SURVIE	9	41	50
	18.00	82.00	100.00
Total	53	48	101
	52.48	47.52	100.00

Il faut bien comprendre que ces façons de faire ne servent qu'à estimer la capacité prédictive du modèle en l'absence de nouvelles données, mais que les paramètres doivent toujours être estimés à l'aide de la totalité des observations.

18.7.2 Validité d'un score, courbe ROC, *AUC*

Ce qui suit ne concerne que le cas de deux groupes. On appellera ici score une mesure permettant de noter le risque d'appartenir au groupe 1. Un score n'est pas forcément obtenu par une méthode linéaire, toute méthode permettant de calculer une probabilité d'appartenance convient : une probabilité est un score compris entre 0 et 1.

On commencera par étudier la séparation entre les distributions du score selon les deux groupes comme dans la figure 18.17. Cependant l'outil le plus pertinent est la courbe ROC.

Abréviation de « Receiver Operating Curve », cette courbe résume les performances de toutes les règles de classement que l'on peut obtenir en faisant varier le seuil de

décision. Supposons que le groupe à détecter prioritairement soit celui des scores élevés. La règle de décision se compare à un test d'hypothèse entre $H1$ (population 1) et $H0$ (population 2). Le vocabulaire (positifs, négatifs) est issu de problématiques de détection (signal, dépistage médical) et peut se ramener aux concepts d'erreurs de première et seconde espèces du chapitre 14. On appelle faux positif une observation classée en $G2$ alors qu'elle appartient à $G1$ etc. Si l'on désigne par s le seuil au delà duquel on classe en $G1$, on définit la :

sensibilité comme le % de vrais positifs : $1 - \beta = P(S > s/G1)$

spécificité comme le % de vrais négatifs : $1 - \alpha = P(S < s/G2)$:

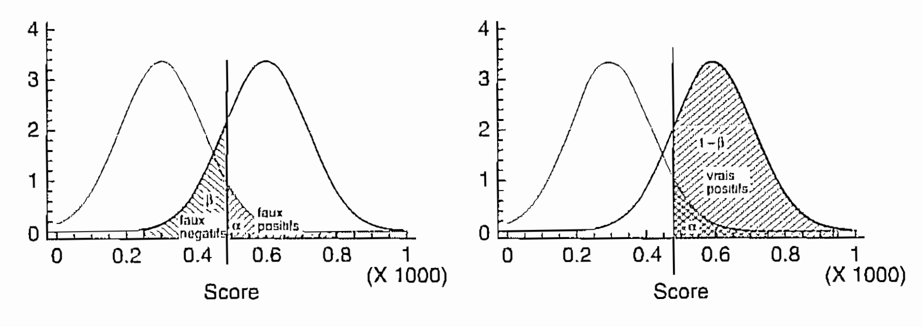


FIGURE 18.19

Si $s = -\infty$ toute observation est classée en $G1$ donc $1 - \beta = 1$ mais $\alpha = 1$. En augmentant s on diminue la sensibilité mais on augmente la spécificité. La courbe ROC (figure 18.20) donne alors l'évolution de la proportion de vrais positifs $1 - \beta$ en fonction de la proportion de faux positifs α.

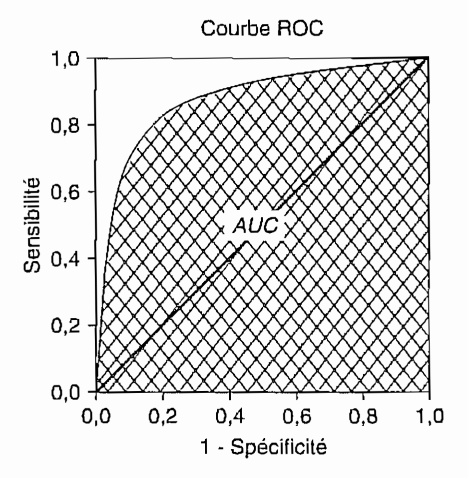


FIGURE 18.20

La courbe ROC est invariante pour toute transformation monotone croissante du score, en raison des propriétés des probabilités : on peut donc sans sans changer la courbe ajouter (ou multiplier par) une constante positive, prendre la probabilité à la place du score etc. La courbe ROC ne dépend que du classement des valeurs.

Plus les deux distributions sont séparées, plus la courbe ROC se rapproche du carré. Si les deux distributions sont identiques, la courbe se confond avec la diagonale. La surface située sous la courbe ROC notée AUC (« area under curve ») est une mesure de la performance d'un score :

$$AUC = \int_{s=+\infty}^{s=-\infty} (1 - \beta(s))d\alpha(s)$$

Elle varie entre 0 et 1, en pratique 0.5 et 1, car si $AUC < 0.5$, cela signifie que les scores ont été inversés. Si $AUC > 0.5$ on utilise également un coefficient dit de Gini qui est le double de la surface comprise entre la courbe ROC et la diagonale et qui vaut donc $2AUC - 1$.

Soit X_1 la variable dont la loi est celle du score conditionnellement à $G1$, idem pour X_2. Un calcul de convolution (loi de $X_1 - X_2$) montre que la surface sous la courbe ROC théorique est égale à $P(X_1 > X_2)$ si l'on tire au hasard et indépendemment une observation de $G1$ et une observation de $G2$.

Cette propriété permet de trouver simplement une estimation de l'AUC. En effet la probabilité que $X_1 > X_2$ s'estime par le pourcentage de paires d'observations (une de $G1$, l'autre de $G2$) concordantes, c'est à dire telles que le score de l'observation de $G1$ est plus grand que le score de l'observation provenant de $G2$. Il y a en tout $n_1 n_2$ paires. La proportion de paires concordantes n'est autre que la statistique U de Mann-Whitney étudiée au paragraphe 14.4.4.2, elle même fonction de la statistique de Wilcoxon.

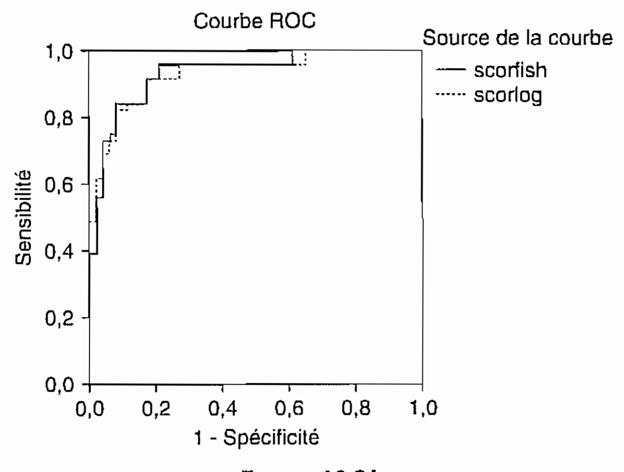


FIGURE 18.21

La figure 18.21 permet de comparer les scores issus de la fonction de Fisher et de la régression logistique pour l'exemple infarctus en ne gardant comme prédicteurs que les variables Incar et Papul : la différence entre les courbes ROC est pratiquement inexistante. Les *AUC* valent respectivement 0.945 et 0.943.

L'*AUC* ne donne toutefois pas un ordre total pour classer des modèles car les courbes ROC peuvent se croiser. De plus quand les modèles ont des complexités différentes, la comparaison doit être effectuée sur les observations d'un échantillon test.

19 Méthodes algorithmiques, choix de modèles et principes d'apprentissage

L'augmentation de la puissance de calcul a permis le développement de nouvelles méthodes de prédiction utilisant une approche algorithmique et issues plus souvent de travaux d'informaticiens que de statisticiens. Arbres de décision, réseaux de neurones, plus proches voisins sont ainsi couramment utilisés en « data mining » et sont en concurrence avec les modèles plus « statistiques » étudiés aux chapitres précédents. D'un autre côté des techniques statistiques élaborées de régression non paramétriques, ou de modélisation non-linéaire via des transformations fonctionnelles (cf. les SVM) sont devenues disponibles.

On peut également combiner différents modèles pour en améliorer les performances : on parle alors de ***méthodes d'ensemble***, de méta-modèles ou méta-heuristiques (« bagging », « boosting » en sont des exemples). Ces approches empiriques peuvent donner lieu à une théorisation expliquant leurs performances ; le lecteur intéressé se reportera à l'excellent livre de T. Hastie, R. Tibshirani, J. Friedman (2001).

La gamme de modèles offerts au praticien est donc de plus en plus vaste.

La question du choix d'un bon modèle, sinon du « vrai modèle », se pose alors en d'autres termes que celui du meilleur ajustement aux données :

- choix d'un modèle parcimonieux utilisant peu de paramètres,
- choix d'un modèle ayant de bonnes capacités prédictives sur de nouvelles observations

Ce chapitre présentera quelques uns de ces aspects ainsi que l'apport de la théorie de l'apprentissage.

19.1 ARBRES DE RÉGRESSION ET DE DISCRIMINATION

Développées autour de 1960 et très utilisées en marketing, ces méthodes délaissées par les statisticiens ont connu un regain d'intérêt avec les travaux de Breiman & al. (1984) qui en ont renouvelé la problématique : elles sont devenues un des outils les plus populaires du ***data mining*** ou ***fouille de données*** en raison de la lisibilité des résultats. On peut les utiliser pour prédire une variable Y quantitative (arbres de régression) ou qualitative (arbres de décision, de classification, de segmentation) à l'aide de prédicteurs quantitatifs ou qualitatifs. Le terme de ***partitionnement récursif*** est parfois utilisé.

19.1.1 Développement d'un arbre binaire

Le procédé consiste à la première étape à diviser l'échantillon d'apprentissage en deux sous ensembles à l'aide d'un des prédicteurs $x^1, x^2, \ldots, x^p$.. Ensuite on recommence séparément dans chaque sous-ensemble etc. Pour chaque variable explicative, il faut donc trouver la meilleure partition de ses valeurs ou modalités en deux sous-ensembles selon un critère d'explication de y.

Il s'agit donc d'une classification descendante à but prédictif opérant par sélection de variables : chaque classe doit être la plus homogène possible vis à vis de y.

Partant de l'ensemble on cherchera à le diviser en deux sous-ensembles d'effectifs n_1 et n_2 tels qu'en moyenne on améliore le plus possible l'homogénéité des deux classes.

Le nombre de divisions en deux sous-ensembles que l'on peut réaliser à l'aide d'un prédicteur (et que l'on doit donc examiner pour choisir la meilleure) dépend de la nature de ce prédicteur :

- si x est qualitatif nominal à m modalités, il y a $2^{m-1} - 1$ dichotomies possibles
- si x est qualitatif ordinal à m modalités et que les coupures doivent respecter l'ordre, il n'y a plus que $m-1$ dichotomies
- si x est numérique à k valeurs distinctes, il y a $k-1$ dichotomies ou coupures possibles entre deux valeurs.

En présence d'un prédicteur qualitatif, on pourrait utiliser des arbres non binaires en découpant en m sous ensembles : cette idée n'est en général pas bonne car elle conduit à des subdivisions avec trop peu d'observations et souvent non pertinentes. L'intérêt des arbres binaires est de pouvoir regrouper les modalités qui ne se distinguent pas vis à vis de y.

19.1.1.1 Arbres de régression

Si y est numérique, on utilisera de façon naturelle la variance de la classe comme mesure d'homogénéité. En divisant en deux sous-groupes on cherche alors à minimiser la variance intra-groupe ou ce qui est équivalent à maximiser la variance inter-groupe. Pour deux groupes la variance inter-groupe $V_{\text{inter}} = \frac{1}{n}(n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2)$ est liée de manière simple à la différence entre les moyennes (calcul laissé au soin du lecteur) : $V_{\text{inter}} = \frac{n_1 n_2}{n^2}(\bar{y}_1 - \bar{y}_2)^2$

La coupure optimale pour une variable qualitative nominale à m modalités doit respecter l'ordre induit par la moyenne de y. On réordonne donc les catégories de x selon $\bar{y}_i$ et il n'y a plus que $m-1$ dichotomies à examiner au lieu de $2^{m-1} - 1$.

19.1.1.2 Discrimination en k classes

Si y est qualitative à m modalités on définit tout d'abord une mesure d'impureté d'un ensemble vis à vis de y. Cette mesure doit être nulle si tous les individus appartiennent à la même modalité de y, maximale si les m catégories sont en proportions égales. Les deux mesures les plus usuelles sont l'***entropie*** $\sum_{i=1}^{k} p_i \ln(p_i)$ et l'***indice de diversité de Gini*** $\sum_{i=1}^{k} p_i(1 - p_i)$. On cherche la division en deux sous-ensembles qui conduit à la diminution maximale de l'impureté.

19.1.1.3 Discrimination en deux classes

Si y n'a que deux modalités en proportions p et $1\text{-}p$ l'indice de Gini vaut $2p(1\text{-}p)$ et a un comportement très proche de l'entropie comme le montre la figure 19.1 où l'entropie a été divisée par $2\ln(0.5)$ pour avoir 0.5 pour maximum.

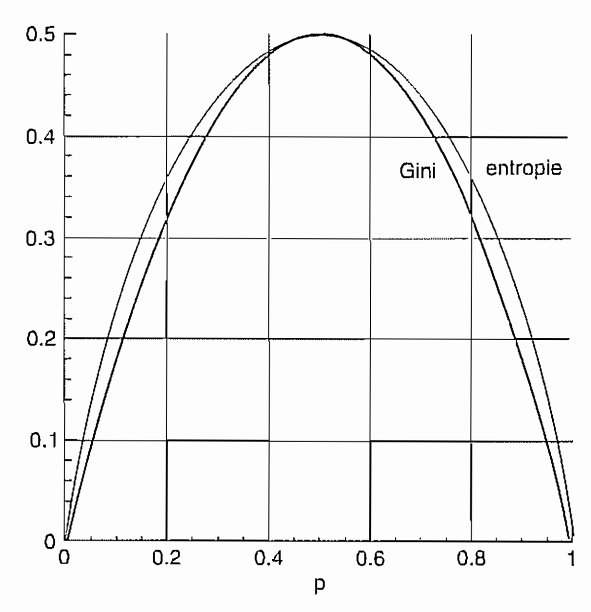


FIGURE 19.1

Sur le plan pratique l'indice de Gini présente les avantages suivants :

– Un calcul simple montre que la diminution d'impureté au sens de Gini vaut $2\frac{n_1 n_2}{n^2}(p_1 - p_2)^2$ où p_1 et p_2 sont les proportions de la modalité 1 dans les deux sous-ensembles obtenus après division. Au facteur 2 près, l'indice de Gini se confond avec la variance de la variable indicatrice de la modalité 1.
– L'indice de Gini présente alors la même propriété que la variance intraclasse qui permet de réduire de $2^{m-1} - 1$ à $m-1$ le nombre de dichotomies à étudier si l'on ordonne les catégories de x selon les proportions d'une des modalités de x.

19.1.2 Utilisation d'un arbre

Pour prédire y, il suffit de parcourir l'arbre depuis le sommet pour déterminer à quel nœud terminal ou segment, appartient une observation $\mathbf{x}$.

Si y est numérique, la prévision sera la moyenne des observations du segment de $\mathbf{x}$. Si les prédicteurs sont numériques, il s'agit d'un modèle de régression constante par morceaux, selon des pavés de $\mathbb{R}^p$ obtenus par dichotomies successives parallèlement aux axes de coordonnées. La qualité de la régression peut être évaluée à l'aide d'indicateurs classiques (erreur quadratique, R^2 etc.)

Si y est qualitative, $\mathbf{x}$ sera classé dans le groupe le plus fréquent (règle majoritaire). On établit alors comme en discrimination un tableau de classement. On peut aussi attribuer à $\mathbf{x}$

une probabilité conditionnelle d'appartenance, à partir des proportions des groupes dans le segment, d'où la possibilité de tracer éventuellement une courbe ROC, mais le nombre de valeurs distinctes de cette probabilité conditionnelle est souvent faible puisque égal au nombre de segments terminaux.

Voici à titre d'exemple (figure 19.2) un arbre obtenu sur les données du paragraphe 18.4.3 avec un sous-échantillon de 374 individus :

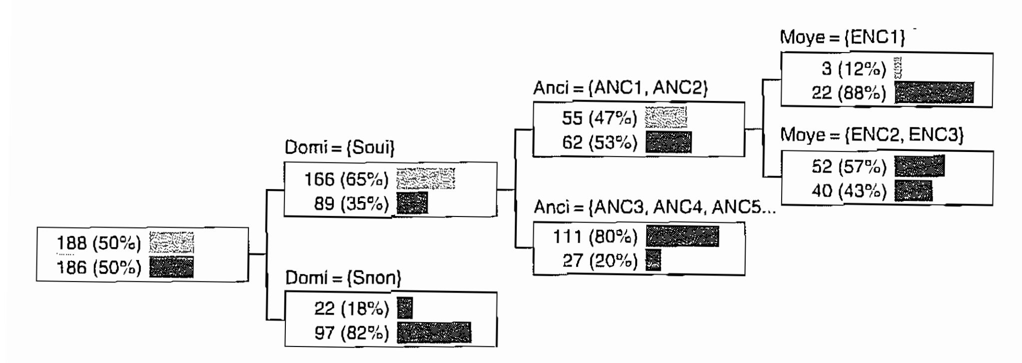


FIGURE 19.2

Cet arbre se lit sous forme de règles logiques : un client qui domicilie son salaire ET qui a une ancienneté de moins de 4 ans ET un encours moyen de plus de 2kF est classé « bon » alors que celui qui ne domicilie pas son salaire est classé « mauvais ». Le taux d'erreur de l'arbre s'obtient en faisant la somme des effectifs des minoritaires de chacun des 4 segments terminaux.

Matrice de confusion

	PREDIT	
OBSERVE	BON	MAUV
BON	163	25
MAUV	67	119

Le taux d'erreur est de 24,6 % du même ordre que celui de la méthode de score mais sur l'échantillon d'apprentissage.

19.1.3 Sélection d'un sous-arbre

Le nombre de nœuds terminaux croit exponentiellement avec le niveau de l'arbre et il est nécessaire de fixer des limites, sinon l'arbre est trop grand et inutilisable car s'ajustant trop bien aux données d'apprentissage : en laissant croître indéfiniment l'arbre il peut se faire que l'on ne s'arrêtera qu'avec des nœuds terminaux réduits à une seule observation. Le taux d'erreur de classement sera alors nul, puisque chaque individu sera affecté à sa classe !

Jusqu'aux travaux de Breiman & al (1984), l'usage était de faire des tests d'égalité de moyennes ou de proportions en se fixant des seuils pour déterminer si un nœud devait être

découpé ou non. Ces tests en cascade conditionnés par les décisions précédentes ont été à juste titre critiqués. La méthodologie « CART » consiste à ne pas fixer de seuil, à laisser croître l'arbre avec pour seul critère d'arrêt un effectif minimal par nœud et ensuite de procéder à un élagage astucieux en utilisant un échantillon test ou une procédure de validation croisée.

Voici succinctement[1] les principes de la méthode dans le cas de la discrimination (prévision d'une variable qualitative).

Soit T_0 l'arbre maximal obtenu comme nous venons de l'indiquer. L'objectif est de trouver un sous-arbre T de T_0 obtenu en coupant certaines branches et qui réalise un bon compromis entre sa performance mesurée par le taux ou coût d'erreur en apprentissage C(T) et sa complexité mesurée par le nombre de segments terminaux $|T|$. On utilise une mesure pénalisée de la performance égale à $C(T) + \alpha\,|T|$ où α est un paramètre de réglage que nous préciserons plus tard.

En termes d'erreur de classement le meilleur arbre est forcément le plus grand. Considérons pour simplifier que $|T_0| = 2^q$ avec q niveaux. Il existe $|T_0|/2$ sous-arbres avec $|T_0| - 1$ segments terminaux obtenus en supprimant une des dernières divisions au niveau $q - 1$. On choisit alors le sous-arbre le meilleur en terme de coûts d'erreur C(T). On poursuit alors l'élagage pour obtenir un sous-arbre à $|T_0| - 2$ segments terminaux etc. jusqu'à arriver à la racine. On dispose alors d'une suite de sous-arbres emboîtés (les branches coupées ne repoussent pas..) de qualité de moins en moins bonne.

La figure 19.3 illustre cette démarche en partant d'un arbre à 8 terminaux (figure 19.3a), il y a 4 sous arbres à 7 terminaux. Le meilleur est celui de la figure 19.3b. Il y a ensuite 3 sous–arbres à 6 terminaux dont le meilleur est en 19.3.c. Il reste ensuite deux choix pour un sous-arbre à 5 terminaux 19.3.d, puis une fois ce choix fait, deux possibilités pour un sous-arbre à 4 terminaux et ensuite il n'y a plus de choix pour passer à 3 (19.3.f) puis 2, puis 1 segment.

Une solution simple pour choisir un de ces sous-arbres consiste à utiliser un échantillon-test et déterminer lequel de ces sous-arbres a la meilleure capacité prédictive, mais on risque de trouver un arbre complexe.

La solution de Breiman & al. est plus élaborée : c'est ici qu'intervient le paramètre α (ce qui précède revient à prendre $\alpha = 0$). Pour α fixé, il existe un sous-arbre minimisant $C(T) + \alpha|T|$ puisque quand $|T|$ diminue C(T) augmente. Pour trouver la valeur adéquate de α on procède par validation croisée : on divise les données disponibles en 10 parties (par exemple), que l'on ôte à tour de rôle et que l'on prédit à l'aide des 9 autres. On fait varier α et on choisit la valeur qui minimise la moyenne des coûts d'erreur.

La méthode s'étend aux arbres de régression en prenant pour C(T) la moyenne des carrés des erreurs.

19.1.4 Avantages et inconvénients

Le principal avantage est l'extrême lisibilité qui fait que tout utilisateur peut comprendre et utiliser un arbre. Parmi les autres avantages figure le fait de pouvoir utiliser des prédicteurs de toute nature, de ne faire aucune hypothèse sur leurs distributions, de hiérarchiser et sélectionner les prédicteurs.

1 Voir Nakache, Confais (2003) pour un traitement détaillé.

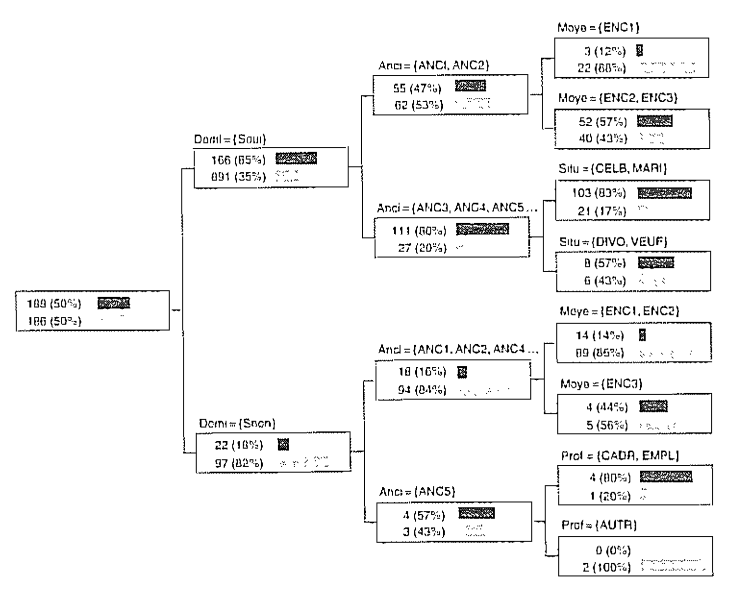


FIGURE 19.3a

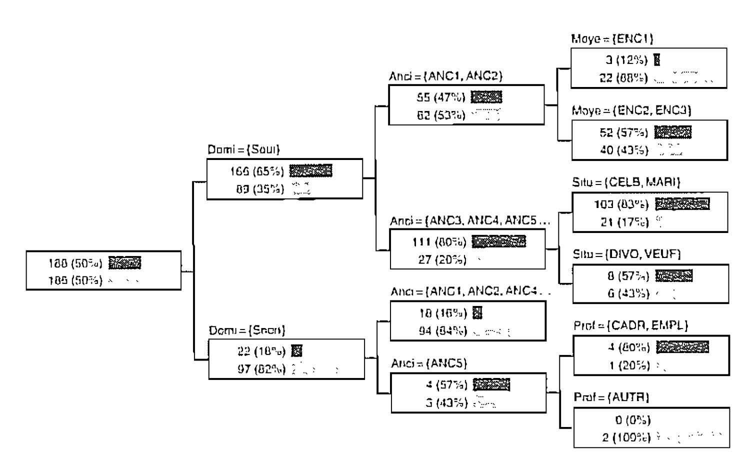


FIGURE 19.3b

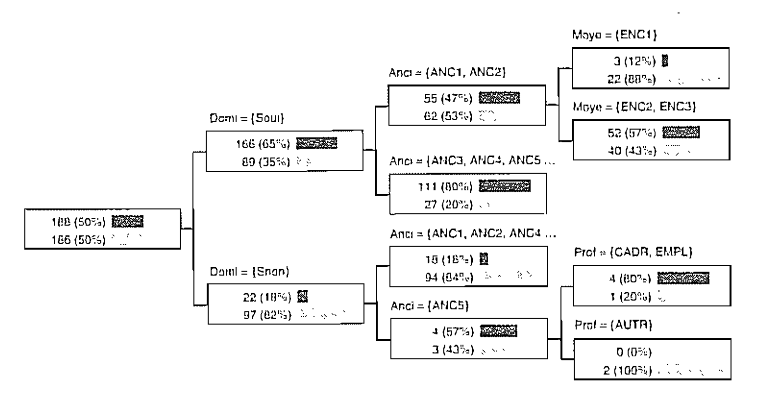


FIGURE 19.3c

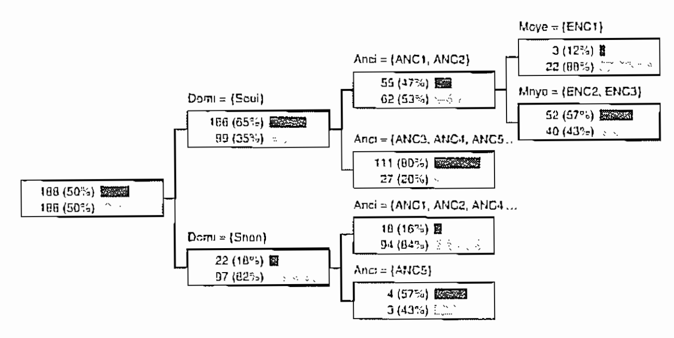


Figure 19.3d

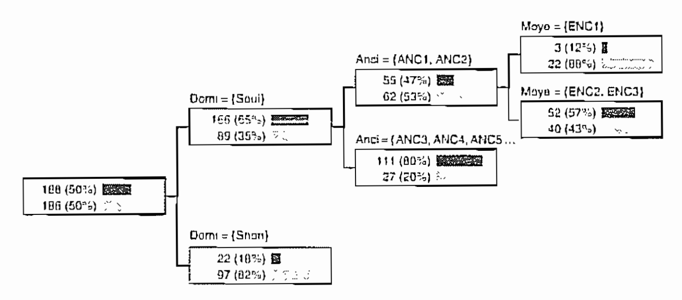


Figure 19.3e

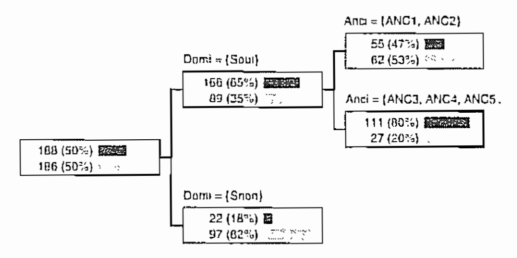


Figure 19.3f

Malgré les progrès méthodologiques les inconvénients sont non négligeables : les arbres sont instables ; de légères variations dans les données peuvent conduire à d'autres choix de nœuds qui deviennent irréversibles puisque chaque coupure détermine les autres. Les arbres ne peuvent être utilisés qu'avec de grands échantillons de plusieurs centaines ou milliers d'observations. C'est pour cela que nous n'avons pas présenté d'arbres pour les données infarctus ou voitures.

19.2 RÉSEAUX DE NEURONES

Les réseaux de neurones sont des outils puissants pour prédire des phénomènes non linéaires. Développés dans les années 80, ils ont connu un vif succès auprès d'utilisateurs non statisticiens cherchant avant tout des performances, grâce en partie à leur vocabulaire

évoquant des analogies biologiques. Ils apparaissaient magiques et mystérieux (Hastie & al. p. 350). La technique peut pourtant se décrire simplement. Nous nous limiterons au cas du réseau le plus connu : le perceptron multicouche, issu des premiers travaux de Rosenblatt (1958). Pour des compléments sur les relations entre « réseaux de neurones et statistique », on se reportera au livre ayant ce titre édité par S. Thiria & al. (1997).

19.2.1 Le perceptron multicouche

Il s'agit d'un modèle utilisable aussi bien en régression qu'en discrimination. Les prédicteurs $x^1, x^2, \ldots, x^p$ sont numériques.

Décrivons un réseau monocouche à k neurones « cachés » :

- On commence par définir plusieurs combinaisons linéaires $w_{0k} + \sum_{j=1}^{p} w_{jk} x_j$ des prédicteurs.

Les coefficients w_{jk} sont parfois appelés *poids synaptiques*. Pour le statisticien ce sont des paramètres à estimer

- Ces combinaisons linéaires sont ensuite transformées par une fonction non-linéaire, dite *fonction d'activation* qui est en général une fonction logistique appelée parfois *sigmoïde* :

$$z_k = \frac{e^{w_{0k} + \sum_{j=1}^{p} w_{jk} x_j}}{1 + e^{w_{0k} + \sum_{j=1}^{p} w_{jk} x_j}}$$

Le *neurone* est le calculateur qui effectue ces opérations et est représenté par un nœud ou une petite boîte sur le schéma.

- Les z_k sont ensuite combinés de façon similaires pour aboutir à des valeurs de sortie qui sont prises pour prévision $\hat{y}$ de y. La sortie est unique pour une régression simple, multiple sinon, comme pour une discrimination. Le ou les $\hat{y}$ sont alors des fonctions non linéaires complexes des $x^1, x^2, \ldots, x^p$.

On retrouve des modèles classiques dans certaines configurations particulières : la régression logistique pour un réseau sans couche cachée avec y binaire.

La figure 19.4 (logiciel Weka) illustre un tel réseau pour une discrimination entre les trois espèces d'iris : les 4 variables alimentent 2 neurones d'une couche cachée dont les sorties sont combinées pour obtenir 3 fonctions, une pour chaque espèce. Une observation est alors classée dans l'espèce qui correspond à la valeur maximale des 3 sorties.

Ce réseau comporte $2 \times 5 + 3 \times 3 = 19$ paramètres à estimer. En effet avec p variables, c neurones sur la couche cachée, et s sorties, il y a $p + 1$ coefficients pour chacune des c fonctions z_k, puis $c + 1$ coefficients pour chaque sortie soit en tout $c(p + 1) + s(c + 1)$ paramètres.

Dans un réseau multicouche les sorties d'une couche deviennent les entrées d'une autre couche etc. On introduit parfois une entrée supplémentaire correspondant à une variable constante égale à 1 pour gérer les termes constants dans les formules.

Le perceptron multicouche (une couche suffit) possède une propriété d'approximation universelle au sens où toute fonction f de p variables $x^1, x^2, \ldots, x^p$ peut être approximée d'aussi près que l'on veut en augmentant le nombre de neurones de la couche cachée (Hornik & al. 1989), à condition d'utiliser une fonction d'activation non linéaire comme la logistique.

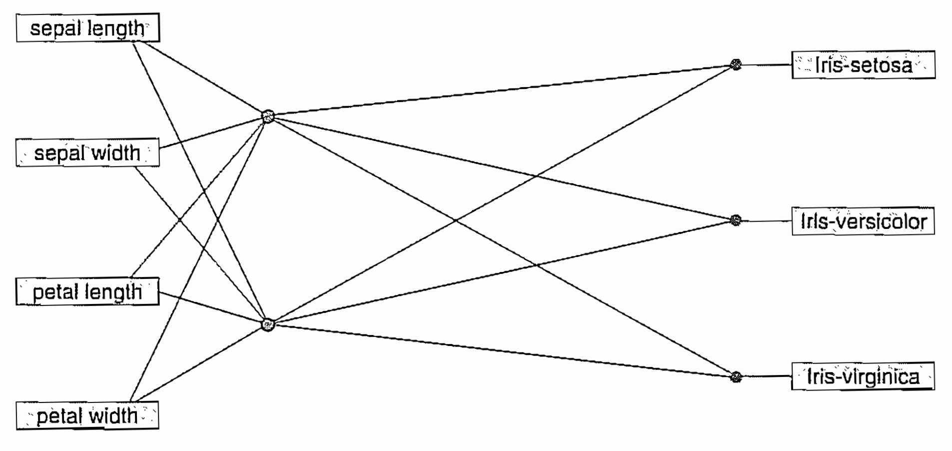


FIGURE 19.4

Les réseaux de neurones avec perte quadratique et fonction d'activation linéaire aboutissent aux mêmes modèles que la régression linéaire ou la discrimination linéaire de Fisher et ne présentent donc pas d'intérêt pratique.

19.2.2 L'estimation

Les paramètres sont estimés pour minimiser une fonction de coût (somme des carrés des écarts si la réponse est numérique, coût d'erreur de classement en discrimination). Compte tenu du caractère non-linéaire, on recourt à des algorithmes d'optimisation numérique que nous ne détaillerons pas ici. Certains algorithmes, de type gradient stochastique, consistent au cours de la phase d'apprentissage à lire plusieurs fois séquentiellement les données en modifiant au fur et à mesure les coefficients pour améliorer la prédiction des valeurs suivantes. Cette phase d'apprentissage peut être extrêmement longue.

Outre le fait que ces algorithmes peuvent aboutir à des optimums locaux, le problème essentiel est le surapprentissage dû au grand nombre de paramètres dès que le réseau est un peu complexe : pour profiter de la propriété d'approximateur universel on prend souvent un nombre élevé de neurones sur la couche cachée. Le choix de l'architecture du réseau : nombre de couches et de neurones par couche est également délicat et se résout par des procédés empiriques comme l'emploi d'ensembles de test ou la validation croisée. Le surapprentissage conduit à des coefficients instables et on emploie alors des méthodes de *régularisation* du type régression ridge, déjà étudiée au chapitre 17 paragraphe 17.5.2 appelée ici « *weight decay* ». On minimisera sur l'échantillon d'apprentissage $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_j (w_j)^2$ où λ est un paramètre de réglage positif. Plus λ est grand plus les coefficients sont réduits, le choix de λ se faisant typiquement par validation croisée.

Les variables d'entrée $x^1, x^2, \ldots, x^p$ doivent au préalable être standardisées afin que la régularisation les traite de la même manière. Les réseaux de neurones sont conçus pour des x_j numériques. Lorsque les prédicteurs sont qualitatifs, on peut utiliser les variables indicatrices des modalités, mais il est préférable de procéder comme dans la méthode Disqual avec les coordonnées sur des axes factoriels.

Bien que la prédiction puisse s'écrire comme une formule mathématique puisque l'on enchaîne des combinaisons linéaires et des fonctions logistiques, cette formule est d'une complexité telle qu'en général elle n'est pas explicitée et le réseau est ensuite utilisé en « boîte noire ». Si l'avantage des réseaux est leur flexibilité, un inconvénient majeur dans certaines applications est l'absence de lisibilité.

Faut-il utiliser les réseaux de neurones ? Ils ont prouvé leur efficacité et font maintenanrt partie de la panoplie des outils disponibles, mais la difficulté à bien les paramétrer est un handicap. D'autres méthodes comme les SVM permettent d'obtenir plus facilement des résultats de qualité comparable avec souvent une formulation plus simple.

19.3 COMBINAISON DE MODÈLES

Parfois appelées méthodes d'ensemble, méta-heuristiques, les méthodes présentées dans ce paragraphe cherchent à améliorer les prédictions obtenues à l'aide de différents modèles par combinaison ou pondération. Il peut s'agir de modèles issus de la même famille (modèles paramétriques ou non) comme des arbres de décision obtenus à l'aide d'échantillons différents, ou de modèles distincts (régression logistique et réseaux de neurones).

La pondération probabiliste de modèles sera étudiée au paragraphe 19.4 lors de l'étude du critère BIC.

19.3.1 Retour sur le bootstrap

L'étude par bootstrap d'une méthode consiste à tirer avec remise B échantillons de taille n dans l'ensemble des n données disponibles. On peut ainsi étudier les distributions d'échantillonnage approchées des paramètres et performances de ces méthodes (voir le paragraphe 18.4.3). Dans le cadre d'une modélisation prédictive, on obtient B modèles différents : appliqués à une observation $\mathbf{x}$, ils fournissent B prédictions différentes de y.

Si y est une variable numérique le *bagging* ou *bootstrap averaging* consiste à prendre la moyenne des B prédictions, qui a donc une variance inférieure à celle de la prédiction initiale et réalise un lissage.

Un cas intéressant est celui où y est qualitative et où on utilise une méthode d'arbres de décision. On se retrouve alors avec B arbres (une forêt !). On procède alors à un vote majoritaire pour classer une observation $\mathbf{x}$: on compte le nombre d'arbres parmi B qui classent $\mathbf{x}$ dans chaque groupe et on choisit le groupe majoritaire. Le bagging remédie à l'instabilité bien connue des arbres, mais la règle finale n'est pas un arbre et on perd la lisibilité de la méthode. De plus on montre que dans le cas des arbres, si le bagging peut améliorer une bonne règle, il n'améliore pas une mauvaise règle mais au contraire l'aggrave (Hastie & al. p. 249) en raison du caractère discontinu de la fonction de perte. Le *boosting* n'a pas cet inconvénient.

19.3.2 Le boosting

Le boosting inventé en 1997 par Freund et Schapire, améliore une règle de discrimination, surtout si elle est médiocre, en l'appliquant de manière répétée sur les observations mal classées en les surpondérant à chaque fois. Le principe consiste donc à se focaliser sur les

observations mal classées, souvent proches de la frontière, plutôt que sur celles faciles à classer. Après chaque itération, on repondère les individus.

La règle finale est un vote pondéré à partir des M règles obtenues : à chaque règle est affecté un coefficient α_m. La procédure AdaBoost se déroule schématiquement de la manière suivante :

- À la première itération les poids des observations sont tous égaux
- A l'itération m on calcule le taux d'erreur pondéré e_m (moyenne des poids des observations mal classées).
- On en déduit le coefficient $\alpha_m = \ln\left(\frac{1 - e_m}{e_m}\right)$
- On met à jour les poids des individus de la façon suivante : si i est bien classé, son poids ne change pas, sinon il est multiplié par $\exp(\alpha_m)$. Quand on normalise pour avoir une somme des poids égale à 1, les poids des observations mal classées augmentent et ceux des biens classés diminuent donc.

Le boosting donne des améliorations spectaculaires pour les arbres. De nombreux travaux ont été mené pour l'expliquer (cf. Hastie et al. chapitre 10). Le défaut est cependant le même que pour le bagging puisque l'on perd l'avantage de la lisibilité de l'arbre. Pour une application donnée il faut alors comparer son efficacité à celles d'autres méthodes de type « boîte noire » comme les réseaux de neurones, la discrimination par estimation de densité, les plus proches voisins etc.

19.4 CHOIX DE MODÈLES

Nous entendrons ici par modèle aussi bien des modèles paramètriques classiques (régression linéaire, logistique) que des méthodes algorithmiques. Devant un ensemble de données, le praticien se trouve alors face au choix d'un modèle parmi un grand nombre de possibilités. Cette question a déjà été évoquée partiellement au chapitre 17 paragraphe 17.4 dans le contexte du choix de variables en régression linéaire multiple.

Depuis les années 1970 où les critères d'Akaïké et de Schwartz ont été proposés, une abondante littérature a été consacrée au choix de modèles et les recherches en ce domaine sont toujours actives. Avant de présenter les principaux critères et méthodes, il faut s'interroger sur l'objectif poursuivi : cherche t-on à découvrir le « vrai » modèle parmi une famille, ou le modèle le plus performant ? La distinction ne va pas de soi et renvoie à des questions épistémologiques. En tout cas le choix de modèle ne sera pas le même.

19.4.1 Critères de vraisemblance pénalisée

On considère ici des modèles paramétrés pouvant se décrire par une densité $g(\mathbf{x} ; \theta)$. Pour un problème prédictif, il pourra s'agir de la densité conditionnelle de y sachant $\mathbf{x}$ ou de la densité conjointe de y et $\mathbf{x}$. Les paramètres seront estimés par la méthode du maximum de vraisemblance.

La vraisemblance calculée en $\hat{\theta}$, $L(\hat{\theta})$, est une manière de mesurer l'adéquation d'un modèle aux données puisqu'elle représente la probabilité d'avoir observé l'échantillon sous le modèle (cf. chapitre 13). On utilisera en fait la log-vraisemblance $\ln L(\hat{\theta})$. Si l'on dispose d'une famille de modèles $g_i(\mathbf{x} ; \theta_i)$ par exemple des régressions linéaires avec 1, 2, ..., p

prédicteurs, on peut calculer pour chaque modèle $\ln L(\hat{\theta}_i)$ mais ce critère ne permet pas de choix car il est croissant avec i : le « meilleur » modèle est celui qui a le plus de paramètres.

Les critères *AIC* et *BIC* vont pénaliser la log-vraisemblance pour tenir compte du nombre de paramètres. D'apparence semblable, ils visent en réalité des objectifs différents.

19.4.1.1 Le critère AIC d'Akaïké

Il vaut

$$\boxed{AIC = -2\ln L(\hat{\theta}) + 2k}$$

où k est le nombre de paramètres du modèle. Le meilleur modèle est donc celui qui minimise *AIC*.

Ce critère tire son origine de la divergence de Kullback-Leibler issue de la théorie de l'information. Soient f et g deux densités de probabilités, et supposons que f est la vraie loi inconnue, g une approximation, alors la divergence, ou perte d'information pour utiliser g à la place de f, est définie par : $I(f\,;g) = \int f(t)\ln\frac{f(t)}{g(t)}dt$. La divergence peut se mettre sous forme de la différence entre deux espérances prises par rapport à la vraie loi:

$$I(f\,;g) = \int \ln(f(t))f(t)dt - \int \ln(g(t))f(t)dt = E_f(\ln(f(t)) - E_f(\ln(g(t))$$

L'élément le plus proche de f dans une famille paramètrée $g(t\,;\theta)$ correspond au θ qui maximise $E_f(\ln(g(t\,;\theta))$. On ne peut résoudre ce problème si f est inconnu. On utilise alors l'estimateur du maximum de vraisemblance $\hat{\theta}$, obtenu dans le cadre de la famille g que l'on porte dans la formule d'où $E_f(\ln(g(t\,;\hat{\theta}))$. Cette dernière expression est une variable aléatoire car $\hat{\theta}$ dépend des données ; on en prend alors l'espérance par rapport aux données (qui suivent la vraie loi f) que l'on note $E_{\hat{\theta}}E_f(\ln(g(t\,;\hat{\theta}))$. Cette quantité n'est pas calculable puisque f est inconnu, mais sous certaines hypothèses et à l'aide d'un développement de Taylor, Akaïké a montré qu'asymptotiquement[2], donc pour de grands échantillons, $E_{\hat{\theta}}E_f(\ln(g(t\,;\hat{\theta})) \sim \ln L(\hat{\theta})) - k$. L'*AIC* s'en déduit par multiplication par -2

19.4.1.2 Le critère BIC de Schwartz

Il vaut

$$\boxed{BIC = -2\ln L(\hat{\theta}) + \ln(n)k},$$

la pénalisation est donc plus forte qu'avec l'*AIC* car dépendant du nombre d'observations. Pour de grands échantillons, le *BIC* aura donc tendance à favoriser des modèles à moins de paramètres que le critère d'Akaïké.

Le critère *BIC* provient d'un contexte totalement différent, celui du choix bayésien de modèles. Considérons une famille finie de m modèles notés M_i dépendant d'un paramètre (vectoriel) θ_i. On se donne des probabilités *a priori* $P(M_i)$ sur chaque modèle, ainsi qu'une

---

2 La démonstration, longue et technique, est omise.

distribution *a priori* de θ_i pour chaque modèle $P(\theta_i/M_i)$; alors la probabilité *a posteriori* du modèle M_i sachant les données **x** est proportionnelle à $P(M_i)\ P(\mathbf{x}/M_i)$

Si les probabilités *a priori* $P(M_i)$ sont uniformes ce qui correspond à ne favoriser aucun modèle, la probabilité *a posteriori* du modèle M_i est proportionnelle à $P(\mathbf{x}/M_i) = \int P(\mathbf{x}/M_i\,;\,\theta_i)P(\theta_i/M_i)d\theta_i$ dite vraisemblance intégrée. Sous certaines hypothèses de régularité, et en effectuant un développement limité au voisinage de l'estimateur du maximum de vraisemblance, on montre (démonstration omise) que $\ln(P(\mathbf{x}/M_i) \sim \ln(P(\mathbf{x}/\hat{\theta}_i, M_i) - \frac{k}{2}\ln(n)$. $\ln(P(\mathbf{x}/\hat{\theta}_i, M_i)$ est la log-vraisemblance du modèle M_i. Le choix du modèle le plus probable M_i *a posteriori* revient à choisir celui qui a le *BIC* minimal.

Ayant calculé tous les *BIC* pour chaque modèle, la probabilité *a posteriori* vaut :

$$P(M_i/\mathbf{x}) = \frac{e^{-0.5\,BIC_i}}{\sum_{j=1}^{m} e^{-0.5\,BIC_j}}$$

On peut alors pondérer les modèles avec ces probabilités, pour en déduire une prédiction moyenne pondérée (*model averaging*).

19.4.1.3 Eléments de comparaison et de réflexion

En régression linéaire multiple, sous les hypothèses habituelles, la log-vraisemblance vaut

$$\ln(L) = -\frac{n}{2}\left[\ln\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}\right) + \ln(2\pi) + 1\right]$$

A nombre fixé de variables, le modèle qui a le meilleur R^2 est aussi celui qui maximise l'*AIC* ou le *BIC*. l'*AIC* ou le *BIC* servent alors à comparer des modèles ayant des nombres de prédicteurs différents.

Sur les données « voitures », l'*AIC* retient un modèle à 2 variables, tandis que le *BIC* ne retient qu'un modèle à une variable.

Si le « vrai » modèle fait partie des m modèles de la famille étudiée, sera t-il choisi ? Si n tend vers l'infini on a pu montrer que la probabilité que le *BIC* choisisse le vrai modèle tend vers 1, ce qui est faux pour l'*AIC*. Par contre l'*AIC* va choisir le modèle qui maximisera la vraisemblance de futures données et réalisera le meilleur compromis biais-variance (voir plus loin). L'*AIC* est donc un critère prédictif tandis que le *BIC* est un critère explicatif. Il faudrait donc choisir le critère selon l'objectif et non les utiliser de façon simultanée.

TABLEAU 19.1

Nombre dans le modèle	R-carré	R carré ajusté	AIC	BIC	Variables du modèle
1	0.6379	0.6153	301.1433	**304.2040**	PUIS
2	0.6866	**0.6448**	**300.5430**	305.1183	PUIS POIDS
3	0.6988	0.6342	301.8305	307.7996	CYL PUIS POIDS
4	0.7018	0.6101	303.6495	310.9014	CYL PUIS LAR POIDS
5	0.7087	0.5874	305.2253	314.0329	CYL PUIS LAR POIDS VITESSE
6	0.7091	0.5504	307.2033	317.3025	CYL PUIS LON LAR POIDS VITESSE

Pour n fini des simulations ont montré des résultats contradictoires et le BIC ne choisit pas toujours le modèle dont les données sont issues car il a tendance à choisi des modèles trop simples en raison de sa plus forte pénalisation.

Il nous semble cependant, que malgré leur intérêt intellectuel, ces critères ne sont pas adaptés à de nombreux problèmes concrets. Ils ne s'appliquent bien que dans des contextes correspondant à une maximisation de vraisemblance et pour certains types de modèles (erreurs gaussiennes par exemple) et de méthodes. On ne peut pas aisément les utiliser pour des réseaux de neurones, des modèles non-linéaires ou à variables qualitatives. Le nombre de paramètres ne traduit pas nécessairement la complexité d'un modèle, nous y reviendrons plus loin. Une régression linéaire multiple à p variables correspond à $k = p + 1$, mais si l'on procède à une régularisation de type ridge la complexité est inférieure. Il faudrait alors remplacer k par un « nombre équivalent de paramètres » ce qui n'est pas simple.

Enfin, la notion de « vrai » modèle qui est implicite ou explicite dans ces critères a t-elle un sens ? Un modèle n'est qu'une simplification de la réalité destinée à la faire comprendre et à obtenir des prévisions convenables. George Box, un des plus grands statisticiens contemporains aimait à rappeler que « tous les modèles sont faux : certains sont utiles ». Lorsque le nombre d'observations est grand, les modèles usuels sont en général trop simples pour la complexité du monde réel et donc rejetés par les tests d'adéquation. Que penser alors de l'intérêt des propriétés asymptotiques ?

19.4.2 Approche empirique

19.4.2.1 Le dilemme biais-variance

On peut généraliser aisément les résultats du paragraphe 17.2.3 à un modèle de prédiction du type $y = f(\mathbf{x}) + \varepsilon$. On estime f par $\hat{f}$ à l'aide d'un échantillon et on cherche à prédire une valeur future en $\mathbf{x}_0$. L'erreur de prédiction est $y_0 - \hat{y}_0 = f(x_0) + \varepsilon - \hat{f}(x_0)$. Elle est aléatoire à deux titres d'une part parce que le phénomène n'est pas déterministe à cause de ε et d'autre part parce que la prédiction $\hat{y}_0 = \hat{f}(\mathbf{x}_0)$ est aléatoire : $\hat{f}$ est une estimation plus ou moins précise. L'erreur quadratique moyenne de prédiction est :

$$E(y_0 - \hat{y}_0)^2 = \sigma^2 + E(f(x_0) - \hat{f}(x_0))^2 = \sigma^2 + (E(\hat{f}(x_0)) - f(x_0)^2) + V(\hat{f}(x_0))$$

le premier terme est irréductible, le deuxième représente le carré du biais du modèle (différence entre l'espérance de la prévision et la valeur moyenne de y_0), le troisième la variance de la prédiction.

Plus un modèle sera complexe plus le biais sera faible, mais en général au détriment de la variance qui va augmenter. Le terme de biais correspond à l'ajustement du modèle sur les données dites d'apprentissage, ajustement qui s'améliore avec la complexité du modèle. La variance correspond à la variabilité de la prédiction pour de nouvelles données.

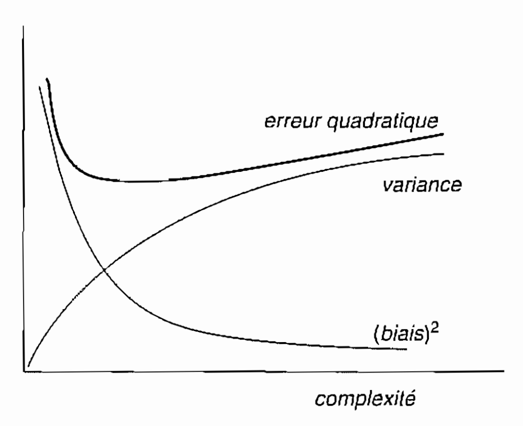


FIGURE 19.5

19.4.2.2 Evaluation et choix de modèle

La figure 19.5 montre qu'il existe un compromis entre biais et variance correspondant à un optimum. Comment l'obtenir empiriquement ? Il faut pour cela estimer l'erreur du modèle sur des données qui n'ont pas servi à l'apprentissage. Lorsque l'on dispose d'un nombre important d'observations, on partagera les données en plusieurs sous-ensembles :

- l'ensemble d'***apprentissage*** sert à estimer chaque modèle en compétition
- l'ensemble de ***validation*** sert à choisir le meilleur modèle, celui qui réalise les meilleures prédictions.
- L'ensemble de **test** sert uniquement à estimer la performance du modèle retenu

On peut ainsi choisir le « bon » modèle quelque soit sa nature, par exemple en faisant varier un paramètre de sensibilité, le nombre de neurones, le nombre de prédicteurs etc.

Par rapport au chapitre précédent paragraphe 18.7.1, on voit qu'un troisième ensemble a été introduit : en effet si l'on doit choisir un modèle en utilisant l'échantillon-test celui ci sert à apprendre le choix de modèle et devient en quelque sorte un échantillon d'apprentissage. On ne peut utiliser alors la mesure d'erreur car elle est biaisée ; il est donc nécessaire de garder des données qui ne servent à rien d'autre qu'à évaluer l'erreur.

Si les données sont en nombre insuffisant, on utilisera la technique de validation croisée qui consiste à partager les données en K sous-ensembles disjoints de même taille et à calculer l'erreur de prédiction moyenne sur chacun de ces sous-ensembles, les K-1 autres formant l'échantillon l'apprentissage. Pour $K = n$ on retrouve la méthode utilisée en analyse discriminante. Le choix de K est encore un compromis biais-variance : K trop grand va donner une grande variance avec un faible biais, tandis que K faible sous-estimera le biais. En pratique $K = 10$ est souvent préconisé.

19.5 LES APPORTS DE LA THÉORIE STATISTIQUE DE L'APPRENTISSAGE DE V. VAPNIK

La théorie développée par V. Vapnik (1998) apporte des vues éclairantes sur ce que l'on appelle la *généralisation* qui n'est autre que la faculté d'un modèle à prédire correctement de nouvelles valeurs et pas seulement à rendre compte du passé. Un grand nombre de résultats font appel à une mesure spécifique de la complexité d'un modèle, la dimension de Vapnik-Cervonenkis, ou VC-dimension notée h. Nous donnons ici un aperçu sans démonstration de cette théorie.

19.5.1 Risque et risque empirique

Soit un modèle de prévision $\hat{y} = f(x\,;\theta)$, où f appartient à une classe paramétrée. On définit alors une fonction de perte $L(y\,;\hat{y})$, en général quadratique, mesurant l'erreur de prévision :

– Si y est numérique, $L(y\,;\hat{y}) = (y - \hat{y})^2$

– Si y est qualitative à deux modalités, L vaut 0 ou 1 selon que l'observation est bien ou mal classée. En prenant y et $\hat{y}$ à valeurs dans $\{-1\,;+1\}$ L peut s'écrire

$$L(y\,;\hat{y}) = \frac{1}{2}\,|\,y - \hat{y}\,| = \frac{1}{2}(y - \hat{y})^2$$

L dépend du paramètre θ.

Le risque est alors l'espérance de la fonction de perte $R = E(L) = \int L(z, \theta)dP(z)$ où $P(z)$ est la loi de probabilité conjointe de y et de $\mathbf{x}$. Le choix optimal de θ serait celui qui minimise R mais c'est une opération impossible quand on ne connaît pas la loi de probabilité $P(z)$.

La méthode courante (moindres carrés par exemple) consiste alors à estimer θ par la valeur $\hat{\theta}$ qui minimise le risque empirique $R_{emp} = \frac{1}{n}\sum_{i=1}^{n} L(y_i\,;f(x_i\,;\theta))$ sur un échantillon (apprentissage) tiré de la loi $P(z)$. Avec d'autres formes de L, on retrouve l'estimateur du maximum de vraisemblance, les estimateurs de Huber etc. R_{emp} est alors une variable aléatoire et on doit se poser la question de sa convergence vers R lorsque n tend vers l'infini pour savoir si la méthode est « consistante ». Pour un modèle donné, le risque empirique est nul si la taille de l'échantillon est trop petite (modèle surparamétré) et croît ensuite jusqu'à atteindre une limite (quand les lois des grands nombres s'appliquent). De son côté, R diminue jusqu'à une valeur limite. Ces deux limites coïncident-elles ? Si elles ne coïncident pas (figure 19.6 à droite), on a un modèle ou processus d'apprentissage non consistant ce qui peut être gênant : en augmentant n on aura une erreur systématique (biais) dans l'estimation de R.

A quelle condition a t-on la consistance ? Paradoxalement cette question s'était peu posée avant les travaux de Vapnik.

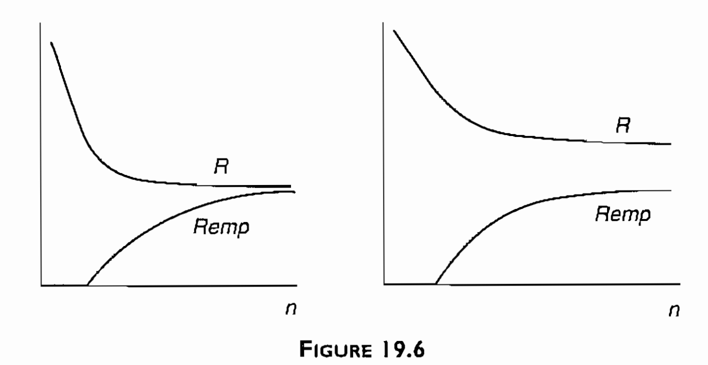


FIGURE 19.6

En moyenne le risque R est toujours supérieur au risque empirique. Avec des modèles surparamétrés, le risque empirique est faible et R grand. Un modèle sera dit *robuste* si les deux risques sont peu différents. Il est facile de trouver des modèles très robustes : le modèle constant $\hat{y} = f(x\,;\theta) = a$ est très robuste mais sans intérêt. Il faut donc réaliser un compromis entre robustesse et ajustement.

19.5.2 La VC-dimension et l'inégalité de Vapnik

Nous nous limiterons maintenant au cas de la discrimination entre deux classes. La dimension de Vapnik-Cervonenkis d'une famille de fonctions de classement (ou *classifieurs*), est une mesure du pouvoir séparateur de cette classe. Ainsi les droites du plan peuvent séparer parfaitement 3 points non alignés (deux d'un groupe, un de l'autre) mais il existe des configurations de 4 points non séparables comme le montre la figure 19.7. La VC-dimension des droites du plan est donc $h = 3$.

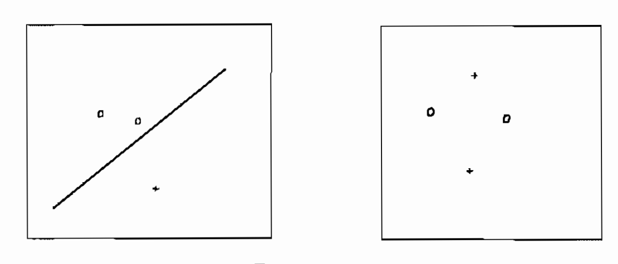


FIGURE 19.7

DÉFINITION :

La VC-dimension d'une famille de classifieurs est le nombre maximal h de points qui peuvent être toujours séparés par la famille de fonctions dans les 2^h configurations où ces points sont libellés ± 1

Cela ne veut pas dire que toute configuration de h points est séparable, ainsi dans le plan, on ne peut pas toujours séparer 3 points alignés, mais que pour $h + 1$ points quelconques il existera toujours une configuration non séparable.

Plus généralement les hyperplans de $\mathbb{R}^p$ ont une VC-dimension égale à $p + 1$. La VC-dimension des paraboles du plan est 4.

La VC-dimension d'une famille de classifieurs n'est cependant pas toujours égale au nombre de paramètres, comme le montre l'exemple suivant classique. Dans $\mathbb{R}$, la VC-dimension des fonctions f définies par $f(x) = 1$ si $\sin(\theta x) > 0$ et $f(x) = -1$ si $\sin(\theta x) < 0$ est infinie car en augmentant θ on peut séparer un nombre arbitraire de points (figure 19.8).

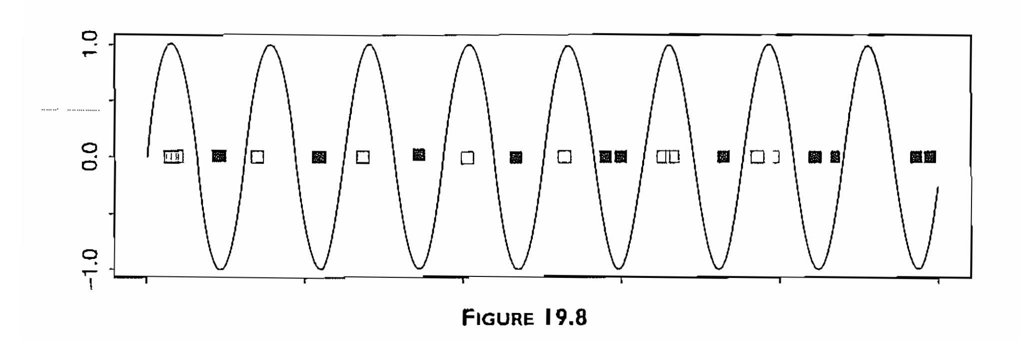


FIGURE 19.8

Revenons maintenant sur les SVM, plus précisément l'hyperplan à vaste marge présenté au 18.3.1.1. Considérons les hyperplans de $\mathbb{R}^p$ dont l'équation est contrainte par $\|\boldsymbol{\beta}\| \leq \frac{1}{C}$, ce qui correspond à une demi-marge supérieure à C (figure 18.12). Soit ρ le rayon de la plus petite sphère contenant toutes les observations alors la VC-dimension h est bornée et est inférieure à $p + 1$:

$$h \leq \min\left[ent\left(\frac{\rho^2}{C^2}\right) ; p\right] + 1$$

ent désignant la partie entière d'un nombre. (cf Burges 1998 pour une démonstration rigoureuse).

La VC-dimension est étroitement liée aux performances d'un processus d'apprentissage, ici une famille de classifieurs.

Vapnik a montré les deux résultats suivants :

- la condition nécessaire et suffisante pour avoir la consistance est que h soit fini.
- Avec une probabilité d'erreur α : $R < R_{\text{emp}} + \sqrt{\frac{h(\ln(2n/h) + 1) - \ln(\alpha/4)}{n}}$

L'inégalité de Vapnik donne une borne pour le risque à partir du risque empirique dépendant de h et de n, mais pas de la distribution des observations. Elle est donc universelle.

Plus h est petit, plus le radical se réduit, ce qui explique les bonnes performances des SVM, et de la régression ridge, même avec un grand nombre de variables, lorsque l'on met des contraintes sur les coefficients.

19.5.3 Le principe de minimisation structurée du risque

La borne de l'inégalité est la somme du risque empirique (l'erreur d'apprentissage) et d'un terme qui ne dépend que du rapport h/n (et de la probabilité d'erreur). On peut donc choisir des modèles plus complexes lorsque le nombre d'observations croît, sans faire augmenter la borne (elle décroit d'ailleurs puisque le risque empirique va décroitre en moyenne en fonction de h).

A n fixé la minimisation de la borne fournit un critère de choix de modèles qui ne fait appel ni à des hypothèses de distributions comme les vraisemblances pénalisées, ni à un échantillon-test : c'est le principe du SRM (*Structural Risk Minimization*). On considère une famille emboîtée de modèles de VC-dimensions croissantes $h_1 < h_2 < \ldots$ (par exemple des modèles linéaires (ou non) à nombre croissant de prédicteurs, des perceptrons multicouches où on augmente le nombre de neurones de la couche cachée, ou le nombre de couches ayant le même nombre de neurones etc.). Pour chaque valeur h_i, on estime le modèle sur les données et on calcule le risque empirique. Le risque empirique décroit (en moyenne) avec h tandis que $\sqrt{\dfrac{h(\ln(2n/h)+1)-\ln(\alpha/4)}{n}}$ croît avec h. On choisit alors le modèle qui correspond au mimum de la somme de ces deux termes : c'est un compromis entre ajustement et robustesse ; la figure19.9 illustre le SRM (rappelons que h est un entier).

L'approche du SRM fournit une solution au problème du choix de modèle en mettant bien en évidence ce qui caractérise véritablement la complexité d'un modèle. Cette approche s'est révélée féconde dans de nombreux cas.

Quelques remarques :

L'inégalité de Vapnik est une inégalité universelle du type des inégalités de Bienaymé-Tchebyshev ou Markov. Elle est intéressante car elle ne dépend pas d'hypothèses sur la distribution des données, en revanche la majoration qu'elle donne peut être très large surtout si h/n est grand : c'est une fonction croissante non bornée de h/n qui peut dépasser 1, ce qui est sans intérêt pour un risque, qui est une probabilité. De nombreux travaux ont été consacrés à la recherche de bornes plus strictes sous certaines hypothèses.

Lorsque la VC-dimension est infinie, l'inégalité ne s'applique pas. On sait qu'il n'y a pas convergence du risque empirique vers R, mais cela n'empêche pas certaines méthodes comme celle du plus proche voisin ou les SVM à noyaux gaussiens, de donner de bons résultats : il y a un biais mais R peut-être faible.

L'inégalité donne une borne avec une probabilité d'erreur, elle n'est donc pas certaine et on peut donc trouver des résultats meilleurs ou pires.

Le calcul de la VC-dimension n'est pas simple et dans bien des cas, on ne connaît pas la valeur de h mais seulement des approximations ou des bornes, ce qui limite l'usage du SRM. Dans le cas où l'inégalité de Vapnik est inutilisable, il vaut mieux choisir le modèle avec une technique de validation croisée.

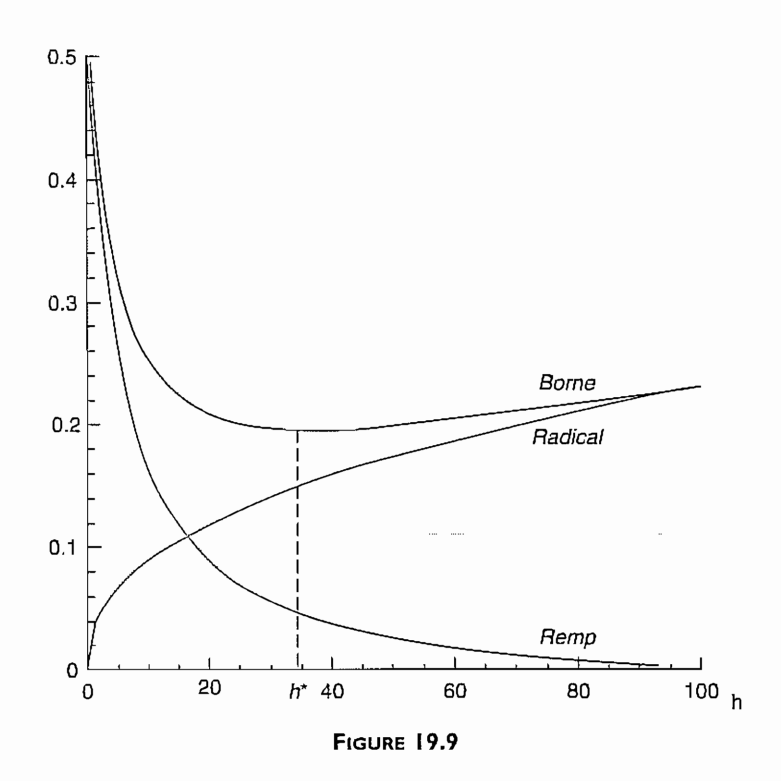


FIGURE 19.9

19.6 PRÉDIRE OU COMPRENDRE ?

La complexité de certains algorithmes de prédiction en font souvent des boîtes noires que l'on ne peut en général pas interpréter. La notion de modèle diffère alors du sens communément établi : il ne s'agit plus d'une représentation de la réalité issue d'une théorie scientifique (physique, économique, biologique, psychologie, . . .) mais seulement d'une technique de prévision. Si le problème est uniquement de prédire, une méthode doit être jugée du point de vue de son efficacité et de sa robustesse : les techniques de choix de modèles de type SRM ou validation croisée apportent une solution. Peut-on prédire sans comprendre ? Cette question peut choquer, mais au delà du débat philosophique, les progrès de outils de calcul semblent bien montrer que oui.

De nombreuses applications ne nécessitent pas de disposer d'une théorie, qui serait d'ailleurs bien difficile à élaborer : par exemple la prévision du comportement des emprunteurs, la détection de segments de consommateurs potentiels d'un produit. La statistique est dans ce cas un outil d'aide à la décision et non un élément de la recherche scientifique.

La « meilleure méthode » est certes celle qui donne les meilleures prévisions, encore faut-il qu'elle soit acceptable lorsqu'elle aboutit à prendre des décisions concernant des personnes. La personne lésée, ou qui pense l'être, est en droit de demander des explications lui

permettant de comprendre la décision[3]. Un arbre de décision, à la rigueur un score linéaire, seront compris, car on pourra expliquer que telle variable a eu telle influence ; il n'en sera pas de même pour une technique d'estimation de densité, un SVM non-linéaire ou un réseau de neurones.

L'acceptabilité des méthodes varie dans le temps, et telle technique qui pouvait paraître complexe à une époque peut devenir usuelle 20 ans plus tard, par suite de la diffusion et de la formation aux outils.

3 En France, la Commission Nationale Informatique et Libertés s'assure que les traitements statistiques ne peuvent nuire aux droits de l'homme et aux libertés individuelles. Voir S. Tuffery (2005).

20 Sondages

20.1 OBJECTIFS ET NOTATIONS

20.1.1 Généralités

Les méthodes de sondage ont pour objectif de tirer dans une population concrète des échantillons destinés à estimer avec la meilleure précision possible des paramètres d'intérêt. Le tirage équiprobable avec remise qui conduit à des échantillons de variables aléatoires indépendantes et identiquement distribuées est la base des développements des chapitres précédents et est le modèle de la statistique mathématique ; ce mode de tirage ne correspond en fait pas à la pratique et n'est au mieux qu'une approximation commode. Les sondages réels portent sur des populations finies et sont effectués par tirage sans remise, pour ne risquer d'interroger deux fois le même individu. Les échantillons ne sont plus constitués de variables indépendantes, et le tirage ne se fait pas toujours avec les mêmes probabilités.

Ce chapitre a pour objectif de donner une initiation à la théorie des sondages aléatoires, et ne prétend nullement couvrir le sujet. En particulier, il faut savoir que les erreurs dues à l'échantillonnage ne sont qu'une partie (pas toujours la plus importante) de l'erreur globale qui comprend les erreurs de couverture, de mesure, de non réponse etc. Bien des sondages sont effectués avec des méthodes non-aléatoires comme la méthodes des quotas qui ne sera pas traitée ici. Le lecteur qui voudrait compléter son information se reportera au livre de P. Ardilly (2006).

20.1.2 Notations

Introduisons maintenant les notations utilisées :

N est la taille de la population. N sera supposé connu, ce qui n'est pas toujours vrai ...

Chaque individu de la population (la population est aussi appelée ***base de sondage***) sera désigné par un identifiant i. On notera Y la variable d'intérêt dont les valeurs sont $(Y_1, Y_2 \ldots\ldots Y_N)$. Y n'est pas une variable aléatoire. On suppose que Y_i sera obtenu sans erreur si l'individu (ou unité) i est sélectionné. Dans ce qui suit Y sera une variable unidimensionnelle numérique, éventuellement binaire quand il s'agira d'estimer des proportions. On s'intéressera à l'estimation de quantités dépendant de Y comme la moyenne $\overline{Y}$ de Y sur la population, ou le total des valeurs $T(Y)$ noté T quand il n'y aura pas d'ambiguïté.

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad T = \sum_{i=1}^{N} Y_i$$

On notera : $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$ la variance et $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{N}{N-1}\sigma^2$ la variance corrigée de Y. Il peut paraître curieux d'utiliser la variance corrigée quand il ne s'agit pas d'un échantillon, mais cela conduit à des formules plus simples.

Un échantillon est un sous-ensemble de n unités de la population. $\tau = \frac{n}{N}$ est le ***taux de sondage***. Il y a C_N^n échantillons distincts possibles, chacun noté s.

Dans un sondage aléatoire chaque unité i de la population a une probabilité de tirage, ou ***probabilité d'inclusion*** π_i bien définie qui ne doit pas être nulle sous peine de ne pouvoir faire des estimations sans biais. On notera que la somme des probabilités d'inclusion vaut (pour des plans de taille fixe) : $\sum_{i=1}^{N}\pi_i = n$ et que π_i est égale à la somme des probabilités des échantillons qui contiennent l'unité i : $\pi_i = \sum_{s(i \in s)} p(s)$. Un ***plan de sondage*** correspond à une distribution de probabilités sur l'ensemble des échantillons.

On utilisera également les probabilités d'inclusion d'ordre 2 : π_{ij} qui donnent la probabilité que les unités i et j appartiennent à l'échantillon.

On appelle variables de ***Cornfield*** les indicatrices δ_i correspondant à la sélection des unités. Ce sont des variables de Bernoulli telles que : $\delta_i = \begin{cases} 1 & \text{si} \quad i \in s \\ 0 & \text{si} \quad i \notin s \end{cases}$

On a :

$$E(\delta_i) = \pi_i$$
$$V(\delta_i) = \pi_i(1 - \pi_i)$$
$$\text{cov}(\delta_i\,;\delta_j) = \pi_{ij} - \pi_i\pi_j$$

On désignera par une lettre minuscule y_i la valeur trouvée dans un échantillon. Cette valeur est donc aléatoire si le tirage de l'unité i est probabiliste.

La moyenne de l'échantillon sera $\overline{y} = \frac{1}{n}\sum_{i \in s} y_i$. Avec les variables de Cornfield, cette moyenne s'écrit : $\overline{y} = \frac{1}{n}\sum_{i=1}^{N} Y_i\delta_i$.

20.2 LE SONDAGE ALÉATOIRE SIMPLE

Il constitue la base des autres méthodes. C'est un tirage équiprobable sans remise : on a donc $\pi_i = \frac{n}{N} = \tau$ et tous les C_N^n échantillons sont équiprobables.

20.2.1 Estimation de la moyenne

La moyenne de l'échantillon est un estimateur sans biais de la moyenne de la population. En effet :

$$E(\overline{y}) = \frac{1}{n}\sum_{i=1}^{N} Y_i E(\delta_i) = \frac{1}{n}\sum_{i=1}^{N} Y_i\pi_i = \frac{1}{n}\sum_{i=1}^{N} Y_i \frac{n}{N} = \overline{Y}$$

Le calcul de la variance est plus complexe car avec un tirage sans remise, les variables de Cornfield ne sont pas indépendantes, mais par raison de symétrie tous les couples auront la même covariance $\pi_{ij} - \pi_i \pi_j = \pi_{ij} - \left(\frac{n}{N}\right)^2$. Calculons la probabilité d'inclusion d'ordre 2 : il y a C_{N-2}^{n-2} échantillons incluant i et j. Comme ils sont tous équiprobables :

$$\pi_{ij} = \frac{C_{N-2}^{n-2}}{C_N^n} = \frac{n(n-1)}{N(N-1)} = \tau\frac{n-1}{N-1}$$

Après quelques calculs simples on trouve que $\text{cov}(\delta_i\,;\,\delta_j) = -\dfrac{\tau(1-\tau)}{N-1}$

Comme : $V(\overline{y}) = V\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{N} Y_i\delta_i\right) = \dfrac{1}{n^2}\left[\displaystyle\sum_{i=1}^{N} Y_i^2 V(\delta_i) + \sum_{i\neq j} Y_iY_j \,\text{cov}(\delta_i\,;\,\delta_j)\right]$

$$= \frac{1}{n^2}\tau(1-\tau)\left[\sum_{i=1}^{N} Y_i^2 - \sum_{i\neq j}\frac{Y_iY_j}{N-1}\right] = \frac{1}{n^2}\tau(1-\tau)\left[\frac{N}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2\right] = \frac{N}{n^2}\tau(1-\tau)S^2$$

On en déduit $V(\overline{y}) = (1-\tau)\dfrac{S^2}{n}$ qui est donc inférieure à la variance du tirage avec remise.

Comme S^2 est inconnue, on l'estime par $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i\in s}(y_i - \overline{y})^2$ dont on peut montrer le caractère sans biais $E(s^2) = S^2$. On en déduit donc l'estimation de la variance de la moyenne $\widehat{V(\overline{y})} = (1-\tau)\dfrac{s^2}{n}$ et un intervalle de confiance approximatif si n est assez grand :

$$\overline{y} - 2s\sqrt{\frac{1-\tau}{n}} < \overline{Y} < \overline{y} + 2s\sqrt{\frac{1-\tau}{n}}$$

L'estimation d'un pourcentage p s'en déduit en considérant que Y est une variable de Bernoulli de paramètre p. Si f est le pourcentage estimé sur l'échantillon, on a :

$$V(f) = (1-\tau)\frac{p(1-p)}{n}\frac{N}{N-1}$$

que l'on estime par : $$\hat{V}(f) = (1-\tau)\frac{p(1-p)}{n-1}$$

En pratique si le taux de sondage est faible (inférieur à 10 %) on a :

$$\hat{V}(f) \simeq \frac{p(1-p)}{n}$$

et on retrouve les résultats du chapitre 13.

20.2.2 Algorithmes de tirage

Une idée élémentaire consiste à tirer des entiers au hasard uniformément répartis entre 0 et N, ce qui peut se faire avec un générateur de nombres aléatoires : on multiplie u par N et

on arrondit à l'entier supérieur. Cette méthode n'est cependant pas utilisée en pratique car elle présente divers défauts : nécessité d'un grand nombre de décimales si N est grand, existence de doublons. On préfère en général des algorithmes séquentiels permettant d'extraire des enregistrements d'un fichier numéroté de 0 à $N-1$, comme le suivant :

On tire un nombre u : si $u \leq \frac{n}{N}$ le premier enregistrement est sélectionné et on recommence pour le deuxième enregistrement en remplaçant n par $n-1$ et N par $N-1$. Si le premier enregistrement n'est pas sélectionné, on tire un autre nombre u et le deuxième enregistrement est sélectionné si $u \leq \frac{n}{N-1}$. Après chaque tirage de nombre au hasard, N diminue d'une unité, tandis que n ne diminue que si une unité est tirée. On continue ainsi jusqu'à l'obtention des n unités.

20.3 SONDAGE À PROBABILITÉS INÉGALES

Le sondage à probabilité égales est utilisable en l'absence de toute information. Supposons par exemple que l'on veuille estimer une production agricole en tirant au sort un certain nombre d'exploitations. Si l'on dispose d'un annuaire donnant les superficies, il est alors préférable d'effectuer ce tirage avec des probabilités proportionnelles à la superficie.

20.3.1 L'estimateur de Horvitz-Thompson

Supposons ici que l'on cherche à estimer le total de la variable d'intérêt $T = \sum_{i=1}^{N} Y_i$. On montre alors que le seul estimateur linéaire sans biais de la forme $\hat{T} = \sum_{i \in s} a_i y_i = \sum_{i=1}^{N} a_i Y_i \delta_i$ est:

$$\boxed{\hat{T} = \sum_{i \in s} \frac{y_i}{\pi_i}}$$

En effet pour que : $E(\hat{T}) = \sum_{i=1}^{N} a_i Y_i E(\delta_i) = \sum_{i=1}^{N} a_i \pi_i Y_i = \sum_{i=1}^{N} Y_i = T$, il faut que $a_i = \frac{1}{\pi_i}$.

Comme les π_i sont inférieurs à 1, on l'appelle aussi estimateur des valeurs dilatées. L'estimateur de la moyenne s'en déduit aisément :

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$$

La variance s'exprime par :

$$V(\hat{T}) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1 - \pi_i) + \sum_{i \neq j}^{N} \sum \frac{Y_i Y_j}{\pi_i \pi_j}(\pi_{ij} - \pi_i \pi_j)$$

que l'on peut mettre sous la forme de Yates-Grundy :

$$V(\hat{T}) = \frac{1}{2} \sum_{i \neq j}^{N} \sum \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 (\pi_i \pi_j - \pi_{ij})$$

lorsque la taille de l'échantillon est fixe.

On en déduit une estimation de la variance :

$$\widehat{V(\hat{T})} = \frac{1}{2}\sum_{i,j\in s}\sum\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}$$

La formule de Yates-Grundy montre que l'on a intérêt à tirer proportionnellement aux valeurs d'une variable auxiliaire X corrélée (positivement!) à Y ce qui est intéressant en cas d'effet taille (chiffre d'affaires, nombre d'employés, bénéfice . . .).

Il peut arriver que certaines unités soient tirées d'office. Ainsi supposons que l'on veuille tirer 3 individus parmi 6 proportionnellement à :

$$x_1 = 300 \quad x_2 = 90 \quad x_3 = 70 \quad x_4 = 50 \quad x_5 = 20 \quad x_6 = 20$$

Les probabilités d'inclusion doivent donc être $\pi_i = \dfrac{nx_i}{\sum_{i=1}^{N} x_i}$

Ce qui donne $\pi_1 = 3\dfrac{300}{550} \geq 1$. La solution est que l'unité 1 soit tirée avec $\pi_1 = 1$ et donc que

$$\pi_2 = 2\frac{90}{250} = 0.72 \quad \pi_3 = 2\frac{70}{250} = 0.56 \quad \pi_4 = 2\frac{50}{250} = 0.4 \quad \pi_5 = \pi_6 = 2\frac{20}{250} = 0.16$$

20.3.2 Le tirage

Le problème est assez compliqué car il y a une infinité de plans de sondages ayant des probabilités d'inclusion d'ordre 1 fixées. Les probabilités d'inclusion d'ordre 2 jouent ici un rôle important : elles devraient être strictement positives et telles que $\pi_{ij} \leq \pi_i\pi_j$ pour pouvoir estimer sans difficulté la variance. Nous renvoyons au livre de Tillé (2001) pour plus de détails.

Une des méthodes les plus utilisées, mais qui peut conduire à des probabilités d'inclusion d'ordre 2 nulles, est le tirage systématique dans les cumuls. Illustrons cette méthode sur l'exemple précédent.

Il reste à tirer 2 unités parmi les unités numérotées de 2 à 6.

On cumule les probabilités d'inclusion, ce qui donne :

$$\pi_2 = 0.72 \quad \pi_2 + \pi_3 = 1.28 \quad \pi_2 + \pi_3 + \pi_4 = 1.68$$
$$\pi_2 + \pi_3 + \pi_4 + \pi_5 = 1.84 \quad \pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6 = 2$$

On tire ensuite un nombre au hasard u compris entre 0 et 1 et on sélectionne les deux individus dont les probabilités cumulées correspondent à u et $u + 1$. Supposons que $u = 0.48$ l'unité 2 est tirée puisque $u < 0.72$ ainsi que l'unité 4 puisque $1.28 < u + 1 < 1.68$. On vérifiera entre autres qu'il est impossible de tirer simultanément les unités 3 et 4.

20.4 STRATIFICATION

La stratification consiste en des tirages séparés effectués dans des sous-populations. Lorsque ces sous-populations sont plus homogènes que la population elle-même, ce qui est généralement le cas, la stratification permet d'obtenir des estimations plus précises qu'un

sondage aléatoire simple de même taille dans toute la population. C'est donc une méthode extrêmement efficace que l'on peut et doit utiliser aussi souvent que possible.

Dans ce qui suit, on supposera que les tirages dans chaque strate sont effectuées selon le sondage aléatoire simple (équiprobable et sans remise).

20.4.1 Formules de base

On notera $N_1, N_2 \dots N_h \dots N_H$ les effectifs des sous-populations ou strates telles que $N = \sum_{h=1}^{H} N_h$.

La moyenne des moyennes de strates $\overline{Y}_1, \overline{Y}_2 \dots \overline{Y}_h \dots \overline{Y}_H$ pondérée par les effectifs redonne la moyenne générale de la population :

$$\overline{Y} = \sum \frac{N_h}{N} \overline{Y}_h$$

La variance de la population se retrouve avec la formule de la ***variance totale*** (voir chapitre 2) où les variances (non-corrigées) de chaque strate sont $\sigma_1^2, \sigma_2^2 \dots \sigma_h^2 \dots \sigma_H^2$:

$$\sigma^2 = \sum \frac{N_h}{N} \sigma_h^2 + \sum \frac{N_h}{N} (\overline{Y}_h - \overline{Y})^2 = \sigma_W^2 + \sigma_B^2$$

σ_W^2 est la variance intra-strates et σ_B^2 la variance inter-strates.

Les tailles des échantillons sont $n_1, n_2, \dots n_h, \dots n_H$ et on dispose des moyennes et variances corrigées de chaque strate : $\overline{y}_1, \overline{y}_2, \dots \overline{y}_h, \dots \overline{y}_H$ et $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \dots \hat{\sigma}_h^2, \dots \hat{\sigma}_H^2$.

Chaque moyenne $\overline{Y}_h$ étant estimée sans biais par $\overline{y}_h$, la moyenne générale est estimée par :

$$\boxed{\hat{\overline{Y}}_{str} = \sum \frac{N_h}{N} \overline{y}_h}$$

qui est l'estimateur de Horvitz-Thompson.

Sa variance se calcule aisément :

$$\boxed{V(\hat{\overline{Y}}_{str}) = \sum \left(\frac{N_h}{N}\right)^2 V(\overline{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h} \frac{N_h - n_h}{N_h - 1} = \frac{1}{N^2} \sum_{h=1}^{H} N_h(N_h - n_h) \frac{S_h^2}{n_h}}$$

20.4.2 Répartition proportionnelle

Ce cas particulier est celui où le taux de sondage est identique d'une strate à l'autre (on parle abusivement d'échantillon représentatif) :

$$\frac{n_h}{n} = \frac{N_h}{N} \Rightarrow \tau_h = \frac{n_h}{N_H} = \frac{n}{N} = \tau$$

L'estimateur stratifié est alors identique à la moyenne usuelle de l'échantillon :

$$\hat{\bar{Y}}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y} = \hat{\bar{Y}}_{prop}$$

La variance se met sous la forme :

$$V(\hat{\bar{Y}}_{prop}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h(N_h - n_h) \frac{S_h^2}{n_h} = \frac{1}{N^2} \sum_{h=1}^{H} \frac{N_h - n_h}{n_h} N_h S_h^2$$

$$= \frac{1}{N^2} \sum_{h=1}^{H} \left(\frac{N_h}{n_h} - 1\right) N_h S_h^2 = \frac{1}{N^2} \sum_{h=1}^{H} \left(\frac{N}{n} - 1\right) N_h S_h^2$$

soit :

$$\boxed{V(\hat{\bar{Y}}_{prop}) = \frac{N - n}{nN} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2}$$

Si le taux de sondage est faible :

$$V(\hat{\bar{Y}}_{prop}) \simeq \frac{N - n}{nN} \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2 = \frac{N - n}{N} \frac{\sigma_w^2}{n}$$

Or $\sigma_w^2 \leq \sigma^2$, donc $V(\hat{\bar{Y}}_{prop}) \leq \dfrac{N - n}{N} \dfrac{S^2}{n}$ si N est grand ($\sigma \simeq S$) qui est la variance de l'estimateur du sondage aléatoire simple.

Avec les mêmes probabilités d'inclusion d'ordre 1, l'échantillon stratifié représentatif est donc plus efficace qu'un échantillon simple de même taille dès que les $\bar{Y}_h$ sont différents.

20.4.3 Répartition optimale

Lorsque les variances de strates sont connues, on peut encore améliorer l'estimateur stratifié et rechercher l'estimateur optimal à n fixé.

Développons $V(\hat{\bar{Y}}_{str}) = \dfrac{1}{N^2} \displaystyle\sum_{h=1}^{H} N_h(N_h - n_h) \dfrac{S_h^2}{n_h} = \dfrac{1}{N^2} \displaystyle\sum_{h=1}^{H} N_h^2 \dfrac{S_h^2}{n_h} - \dfrac{1}{N^2} \displaystyle\sum_{h=1}^{H} N_h S_h^2$

Le deuxième terme ne dépend pas de l'échantillon. On a alors à résoudre le problème suivant : $\min \displaystyle\sum_{h=1}^{H} N_h^2 \dfrac{S_h^2}{n_h}$ sous la contrainte $\displaystyle\sum_{h=1}^{H} n_h = n$

En annulant les dérivées partielles du Lagrangien, $\displaystyle\sum_{h=1}^{H} N_h^2 \dfrac{S_h^2}{n_h} - \lambda \displaystyle\sum_{i=1}^{H} n_h$ par rapport aux effectifs inconnus (oubliant que ce sont des nombres entiers) on trouve la répartition de Neyman qui montre qu'il faut sur-représenter les strates les plus dispersées par rapport à la répartition proportionnelle :

$$\boxed{n_h = n \frac{N_h S_h}{\sum N_h S_h}}$$

Les effectifs doivent être arrondis. Le calcul peut se généraliser en considérant des coûts d'enquête différents par strate et en optimisant à budget fixé.

On recommande souvent de faire beaucoup de strates pour améliorer la variance interclasse, mais le risque est alors d'avoir des tailles d'échantillon trop faibles dans certaines strates.

20.5 SONDAGE EN GRAPPES ET TIRAGE SYSTÉMATIQUE

On appelle grappes M sous-populations d'effectifs $N_1, N_2 \dots N_m \dots N_M$. La méthode consiste alors à tirer m grappes et à sélectionner **tous** les individus de chaque grappe. On connaîtra donc sans erreur le total et la moyenne de chaque grappe. L'intérêt essentiel de ce mode de tirage est son caractère économique, en particulier si les grappes sont des zones géographiques, car on diminue alors fortement les coûts de déplacement en face à face. Cette méthode est très utile lorsque la taille de la population est inconnue, car on n'aura besoin de connaître que la taille des grappes choisies. On ne pourra cependant pas estimer tous les paramètres. La taille de l'échantillon est aléatoire si les grappes ont des effectifs différents.

20.5.1 Tirage de grappes à probabilités inégales

Cherchons à estimer le total $T = \sum_{i=1}^{M} T_i$. L'estimateur de Horvitz-Thompson est $\hat{T} = \sum_{i=1}^{m} \frac{T_i}{\pi_i}$ où les π_i sont les probabilités de tirage des grappes. La variance de cet estimateur est donnée par les formules du paragraphe 20.3.1 où on remplace Y_i par T_i puisque tout revient à un tirage de m totaux parmi M.

L'estimateur de la moyenne est $\hat{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{m} \frac{T_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{m} \frac{N_i \bar{Y}_i}{\pi_i}$ et nécessite la connaissance de N.

Un cas intéressant est celui où les grappes sont tirées avec des ***probabilités proportionnelles à leur effectif*** $\pi_i = m\frac{N_i}{N}$. La taille de l'échantillon est aléatoire d'espérance

$$E(n_s) = E(\sum_{i \in s} N_i) = \sum_{i=1}^{M} N_i E(\delta_i) = \sum_{i=1}^{M} N_i \frac{N_i m}{N} = \frac{m}{N}\sum_{i=1}^{M} N_i^2 .$$

L'estimateur de la moyenne est alors $\hat{\bar{Y}} = \frac{1}{m}\sum_{i=1}^{m} \bar{y}_i$ et sa variance peut être estimée par :

$$\bar{V}(\hat{\bar{Y}}) = \frac{1}{m(m-1)} \sum_{i=1}^{m}\left(1 - m\frac{N_i}{N}\right)\left(\bar{y}_i - \hat{\bar{Y}}\right)^2$$

Une bonne répartition en grappes est caractérisée par des moyennes de grappes peu différentes de la moyenne générale : c'est donc l'inverse de la stratification : ici les grappes doivent être les plus hétérogènes possibles (chacune doit pouvoir représenter la population).

20.5.2 Tirage de grappes à probabilités égales

Cette fois $\pi_i = \dfrac{m}{M}$ d'où $E(n_s) = E\left(\sum_{i \in S} N_i\right) = \sum_{i=1}^{M} N_i \dfrac{m}{M} = \dfrac{Nm}{M}$.

L'estimateur du total $\hat{T} = \dfrac{M}{m} \sum_{i \in s} T_i$ a pour variance $V(\hat{T}) = \dfrac{M - m}{M - 1} \dfrac{M}{m} \sum_{i=1}^{M} \left(T_i - \dfrac{T}{M} \right)^2$.

L'estimateur de la moyenne est $\hat{\bar{Y}} = \dfrac{M}{m} \dfrac{1}{N} \sum_{i=1}^{m} N_i \bar{Y}_i$.

20.5.3 Le tirage systématique

Ce mode de tirage est très utilisé quand on ne sait pas, ou que l'on ne veut pas, faire un tirage aléatoire équiprobable. Supposons que N est un multiple de n. Par exemple on veut tirer 10 individus parmi 1000 : on commence par tirer au hasard un nombre entier entre 1 et 100, si ce nombre est 27, le premier individu sera le n°27, le deuxième le n°127, etc. jusqu'au n°927. Il s'agit donc en fait d'un tirage d'une seule grappe parmi $M = N/n$ grappes.

De façon générale si l'on a tiré un entier h, les individus sélectionnés ont les numéros : $h, h + M, h + 2M, \ldots, h + (n - 1)M$.

L'estimateur de la moyenne est simplement la moyenne de la grappe sélectionnée et sa variance est $V(\hat{\bar{Y}}) = M \sum_{i=1}^{M} \left(\dfrac{\bar{Y}_i N_i}{N} - \dfrac{\bar{Y}}{M} \right)^2$.

Lorsque le fichier se trouve être trié selon un ordre proche de Y, la variance peut être notablement plus faible que pour le tirage aléatoire simple. On pourra s'en convaincre en prenant à titre d'exercice le cas $Y_i = i$. Il est incorrect d'utiliser la variance de l'estimateur du tirage aléatoire simple sauf si la base de sondage a été préalablement triée au hasard.

20.6 REDRESSEMENT

Lorsque l'on dispose *a posteriori* d'une information supplémentaire corrélée avec la variable d'intérêt Y, on peut améliorer la précision des estimations. Cette information peut être qualitative ou quantitative. Nous exposerons brièvement les principaux cas dans le cadre d'un sondage aléatoire simple, pour une variable d'intérêt quantitative.

20.6.1 Quotient, régression

Le cas suivant est inspiré de : Ardilly, Tillé (2003) page 173. On effectue un sondage auprès de $n = 100$ entreprises parmi $N = 10\,000$ pour estimer le chiffre d'affaires moyen. On trouve $\bar{y} = 5.2\ 10^6$ €. On sait par ailleurs que le nombre moyen de salariés de la population est $\bar{X} = 50$. Or dans l'échantillon on a $\bar{x} = 45$. Comme on soupçonne une relation de proportionnalité entre Y et X, on effectue une règle de 3 : c'est l'estimation par la méthode du quotient :

$$\bar{y}_q = \bar{y} \frac{\bar{X}}{\bar{x}}$$

On corrige donc l'estimation initiale et on trouve $\bar{y}_q \simeq 5.8\ 10^6$ €

Cet estimateur est biaisé, mais le biais est faible si n est grand. Comme il est biaisé, il vaut mieux calculer son erreur quadratique plutôt que la variance. On montre qu'elle est approximativement égale à :

$$E(\overline{y}_q - \overline{Y})^2 = \frac{N-n}{Nn}\left(S_y^2 - 2\frac{\overline{Y}}{\overline{X}}S_{xy} + \left(\frac{\overline{Y}}{\overline{X}}\right)^2 S_x^2\right) \text{ que l'on estime par } \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n} z_i^2$$

avec :
$$z_i = y_i - rx_i \quad \text{où} \quad r = \frac{\overline{y}}{\overline{x}}.$$

Il y a amélioration si $\dfrac{S_{xy}}{S_x^2} > \dfrac{r}{2}$

La méthode du quotient suppose une stricte proportionnalité. Si la relation est du type $Y = a + bX$, il vaut mieux effectuer une régression linéaire et utiliser l'estimateur :

$$\overline{y}_r = \overline{y} + b(\overline{X} - \overline{x})$$

mais pour calculer b, il faut alors disposer des valeurs de X pour chaque unité sélectionnée et pas seulement de la valeur moyenne.

20.6.2 Post-stratification

Lorsque le caractère auxiliaire est qualitatif, l'idée consiste à effectuer un calcul comme pour l'estimateur stratifié vu plus haut :

$$\hat{\overline{Y}}_{post} = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_h$$

en répartissant *a posteriori* les observations selon les modalités du caractère auxiliaire.

La différence essentielle ici est que les effectifs n_h par strates ne sont plus fixés *a priori* mais sont aléatoires de loi hypergéométrique.

L'estimateur reste sans biais (si les post-strates ne sont pas vides), mais sa variance va prendre en compte les fluctuations des n_h.

Le calcul de la variance est assez complexe. On commence par écrire la formule de la variance totale en conditionnant par les n_h et les supposant non-nuls.

$$V\left(\hat{\overline{Y}}_{post}\right) = \underbrace{V\left[E\left(\hat{\overline{Y}}/n_h\right)\right]}_{0} + E\left[V\left(\hat{\overline{Y}}/n_h\right)\right]$$

Le premier terme est nul car l'espérance conditionnelle vaut toujours $\overline{Y}$. La variance conditionnelle vaut :

$$\sum\left(\frac{N_h}{N}\right)^2 V(\overline{y}_h) = \sum\left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h n_h} S_h^2 = \sum\left(\frac{N_h}{N}\right)^2 S_h^2 \frac{1}{n_h} - \frac{1}{N}\sum\left(\frac{N_h}{N}\right) S_h^2$$

par la formule habituelle.

Il faut en prendre ensuite l'espérance $\sum\left(\frac{N_h}{N}\right)^2 S_h^2 E\left(\frac{1}{n_h}\right) - \frac{1}{N}\sum\left(\frac{N_h}{N}\right)S_h^2$. Or il n'y a pas de formule simple pour l'espérance de l'inverse d'une hypergéométrique. Après des développements limités pour n grand, que l'on omettra ici, on trouve finalement :

$$V\left(\hat{\bar{Y}}_{post}\right) \simeq \left(\frac{1-\tau}{n}\right)\sum\frac{N_h}{N}S_h^2 + \frac{1-\tau}{n^2}\sum\left(1-\frac{N_h}{N}\right)S_h^2$$

Le premier terme n'est autre que la variance de la stratification *a priori* avec répartition proportionnelle, ce qui prouve que stratifier *a priori* est toujours meilleur qu'*a posteriori*. Pour que la stratification *a posteriori* soit plus efficace que le sondage aléatoire simple, il faut que le deuxième terme ne soit pas trop grand : cela se produit si le rapport de corrélation $\eta^2(Y/X)$ est grand. Lorsque ce rapport est nul, la stratification *a posteriori* est au contraire moins efficace que le sondage aléatoire simple.

20.6.3 Poids de redressement

Considérons une post-stratification selon H post-strates. L'estimateur de la moyenne de la variable d'intérêt est :

$$\sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h = \sum_{h=1}^{H}\frac{N_h}{N}\frac{1}{n_h}\sum_{I_h}^{n_h} y_{I_h} = \sum_{h=1}^{H}\sum_{I_h}^{n_h}\frac{N_h}{Nn_h}y_{I_h}$$

On appelle poids de redressement le coefficient $\frac{N_h}{Nn_h}$. La somme des poids de redressement sur les n unités de l'échantillon vaut alors 1. Ceci permet d'obtenir l'estimation de $\bar{Y}$ comme une moyenne pondérée des valeurs observées. Il ne faut pas confondre les poids de redressement avec les poids d'échantillonnage (probabilités d'inclusion).

Le redressement consiste à modifier les proportions des post-strates $\frac{n_h}{n}$ pour les rendre égales à $\frac{N_h}{N}$ à l'aide d'une règle de 3.

Lorsque l'on veut redresser sur plusieurs variables qualitatives à la fois (par exemple : sexe, CSP, etc.) le calcul des poids de redressement est plus complexe et s'effectue à l'aide d'algorithmes itératifs dont le plus connu est celui de Deming et Stephan qui consiste en une suite de règles de 3 sur chaque critère.

Exemple : 1 000 individus ont été interrogés. La répartition par sexe et profession est la suivante

	P1	*P2*	*P3*	*Total*
H	300	100	200	*600*
F	100	150	150	*400*
Total	*400*	*250*	*150*	*1000*

Supposons que les vraies marges soient 500 et 500 pour le sexe et 350,300, 350 pour la profession.

Une première règle de 3 permet d'obtenir les marges souhaitées pour le sexe : on multiplie la première ligne par 500/600 et la deuxième ligne par 500/400

	P1	*P2*	*P3*	*Total*
H	250	83	167	*500*
F	125	187.5	187.5	*500*
Total	*375*	*270.5*	*354.5*	*1000*

On redresse ensuite en colonne pour ajuster les effectifs marginaux de la variable profession, ce qui change les marges en ligne :

	P1	*P2*	*P3*	*Total*
H	233	92	165	*490*
F	117	208	185	*510*
Total	*350*	*300*	*350*	*1000*

Puis en ligne :

	P1	*P2*	*P3*	*Total*
H	238	94	168	*500*
F	115	204	181	*500*
Total	*353*	*298*	*349*	*1000*

En l'absence de cases vides, l'algorithme converge rapidement et donne les poids de redressement à appliquer à chaque case. Ainsi à la quatrième itération (très proche du résultat souhaité), les 300 individus H et *P1* ont chacun un poids de 0.236. La somme des poids de redressement des 1000 individus vaut 1000.

	P1	*P2*	*P3*	*Total*
H	236	95	168	*499*
F	114	205	182	*501*
Total	*350*	*300*	*350*	*1000*

L'utilisation de redressement sur plusieurs critères doit être effectuée avec précaution pour éviter des poids trop dispersés ; il ne faut redresser que sûr des critères corrélés avec la variable d'intérêt, sinon on n'améliore pas les estimations.

21 Plans d'expériences

Avec les techniques de sondage, les plans d'expériences constituent la deuxième grande méthodologie statistique pour recueillir des données. Il ne s'agit plus ici d'observer des individus existants en allant les chercher dans leur population, mais de provoquer des résultats, ou « réponse », en faisant varier intentionnellement certains « facteurs » dans le but d'étudier le modèle liant la réponse aux facteurs.

Les objectifs sont divers : par exemple déterminer quels sont les facteurs influents, estimer au mieux le modèle, trouver pour quelles valeurs on peut obtenir une valeur optimale de la réponse . . .

Un des grands intérêts des plans d'expériences est de pouvoir réduire le nombre des expériences à effectuer en les choisissant judicieusement d'où des économies parfois considérables. Ainsi avec 10 facteurs à 2 niveaux chacun, au lieu de faire les $2^{10} = 1024$ expériences possibles, un plan de Plackett et Burman en proposera seulement 12 et un factoriel fractionnaire 16. Mais les résultats ne seront valables que si aucune intéraction n'existe entre les facteurs.

La détermination d'un plan d'expériences, et plus généralement d'un dispositif expérimental, ne peut donc se concevoir en dehors du modèle de régression censé représenter la relation entre réponse et facteurs. Tel plan sera adapté à un modèle sans interaction avec effets du premier degré, tel autre pour un modèle du second degré, tel encore pour un modèle à facteurs qualitatifs, etc.

Dans un modèle linéaire $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, il s'agit donc de trouver la matrice $\mathbf{X}$.

La planification des expériences ne date que du XXème siècle : développée tout d'abord en agronomie avec les travaux de Fisher, puis dans diverses branches de la recherche industrielle en particulier en chimie puis en mécanique.

Ce bref chapitre n'est qu'une introduction à ce vaste domaine, dans le cas de modèles linéaires. Nous renvoyons à l'ouvrage collectif édité par J.J. Droesbeke & al. (1997) pour un traitement plus complet.

21.1 INTRODUCTION

21.1.1 Vocabulaire

La terminologie varie fortement d'un domaine d'applications à l'autre et mérite donc d'être précisée.

Tout d'abord la variable y s'appellera la ***réponse***, les variables explicatives x_j des ***facteurs***. Ces facteurs peuvent être qualitatifs (type d'engrais, marque) avec des ***modalités*** ou bien quantitatifs (température, hygromètrie) avec des ***niveaux***. Une ***expérience*** ou ***essai*** ou ***traitement*** sera une combinaison de modalités ou niveaux des facteurs.

Toutes les combinaisons ne sont pas réalisables, ce qui conduit à définir le ***domaine expérimental***, souvent un hypercube pour des facteurs quantitatifs.

En général le nombre d'expériences réalisables sera fini $\prod_{j=1}^{p} m_j$, mais souvent très élevé, même si chaque facteur ne peut prendre que quelques niveaux m_j.

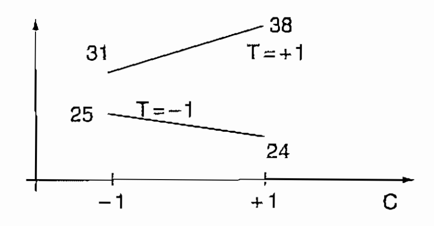
On distinguera la ***matrice d'expériences*** qui est la liste des essais à effectuer, du ***dispositif expérimental*** qui précise l'ordre des essais. Le plus souvent ce essais seront effectués dans un ordre aléatoire obtenu par permutation des lignes de la matrice d'expériences : c'est la ***randomisation***. On recourt également à la mise en ***blocs***, consistant à répartir les essais en sous-ensembles aussi homogènes que possibles. Ces dispositifs ont pour but d'éliminer l'influence de certains facteurs non *contrôlables* comme la température extérieure, l'ensoleillement, etc.

La matrice du modèle **X** se déduit de la matrice d'expériences : on ajoute des colonnes en tenant compte du degré et des interactions entre facteurs. Ainsi pour un modèle linéaire du second degré à 2 facteurs quantitatifs, $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 (x^1)^2 + \beta_4 (x^2)^2 + \beta_5 x^1 x^2 + \varepsilon$. **X** possédera 6 colonnes obtenues en ajoutant une colonne de 1 pour le terme constant, 2 colonnes correspondant aux carrés des variables et une à leur produit.

On parlera d'***effets*** du premier degré, du second degré, d'effets d'***interaction.***

L'interaction entre A et B se traduit par la non additivité des effets au sens suivant : si l'on étudie les variations moyennes de la réponse selon A, l'effet de A ne doit pas dépendre du niveau du facteur B. Illustrons ce concept par l'expérience suivante (adaptée de Sado "Plans d'expériences", AFNOR 1991) : on mesure le rendement Y d'une réaction chimique selon deux facteurs température T et concentration C : T varie de 50 à 100 °C et C varie de 20 à 30 g/l. On recode les niveaux en -1 et $+1$ et on effectue 4 essais aux extrémités du domaine de variation (voir plus loin) :

Essai	T	C	Y
1	−1	−1	25
2	+1	−1	31
3	−1	+1	24
4	+1	+1	38



Le graphique prècédent montre que l'effet de la température n'est pas le même selon le niveau de la concentration : il y a augmentation de Y quand la température augmente, mais cette augmentation dépend de C : elle est de 6 pour $C = -1$ et de 14 pour $C = +1$. L'absence d'intéraction se serait traduite par des segments parallèles.

On verra plus loin que certains plans ne permettent pas d'estimer tous les effets des facteurs, c'est le phénomène de ***confusion*** ou d'***alias***.

21.1.2 Optimalité et orthogonalité

Supposons le nombre n d'essais fixé. S'il s'agit d'estimer au mieux les paramètres du modèle linéaire $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, on cherchera des propriétés d'optimalité pour la matrice de variance-covariance des $\boldsymbol{\beta}$. On sait d'après le paragraphe 17.2.1.1 que cette matrice vaut $V(\beta) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. L'optimum ne dépend que de $\mathbf{X}$ et non de la réponse. Obtenir des estimateurs de variance minimale revient à définir un critère de maximalité pour $\mathbf{X}'\mathbf{X}$.

Le critère le plus utilisé est celui du déterminant maximal ou ***D-optimalité*** $\max|\mathbf{X}'\mathbf{X}|$. Il revient à minimiser le volume de l'ellipsoïde de confiance des β, pour un niveau de confiance donné.

Il existe bien d'autres critères, mais moins utilisés comme la ***A-optimalité*** : $\min(\text{Trace}\,(\mathbf{X}'\mathbf{X})^{-1})$ qui revient à minimiser la somme des variances des estimateurs des β.

La matrice $\mathbf{X}$ doit être de plein rang : pour des facteurs qualitatifs, on éliminera une indicatrice par facteur comme dans le modèle linéaire général.

Pour des facteurs quantitatifs, on a vu au chapitre 17 paragraphe 17.3.2.1 que la variance de chaque coefficient de régression estimé était minimale si les variables explicatives étaient non corrélées deux à deux : les colonnes de $\mathbf{X}$ sont orthogonales. ***Les plans orthogonaux sont donc optimaux***, ce qui a conduit à privilégier leur recherche, d'autant plus que l'interprétation des résultats par l'analyse de variance en est très simple et que les calculs peuvent se faire manuellement, avantage essentiel avant l'apparition des ordinateurs.

Cependant de tels plans n'existent pas toujours : ainsi il est facile de voir qu'il est impossible d'obtenir des colonnes orthogonales pour un modèle linéaire du second degré à cause des termes carrés. On pourra s'intéresser à d'autres propriétés comme l'***isovariance par rotation*** : c'est le cas si la variance de la prédiction de la réponse en un point $\mathbf{x}$, qui vaut $\sigma^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$, ne dépend que de la distance au centre du domaine $(\mathbf{x}'\mathbf{x})^{\frac{1}{2}}$ et est donc indépendante de l'orientation des axes.

21.2 PLANS POUR FACTEURS QUANTITATIFS ET MODÈLE LINÉAIRE DU PREMIER DEGRÉ

Le modèle pour p facteurs est donc celui de la régression linéaire multiple classique : $y = \beta_0 + \beta_1 x^1 + \ldots + \beta_p x^p + \varepsilon$

21.2.1 Le cas de la régression simple

Soit un seul facteur prenant ses valeurs dans un intervalle $[x_{min} ; x_{max}]$. On sait (chapitre 16, paragraphe 16.2.1) que la variance de l'estimateur du coefficient de régression vaut : $V(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Si n est pair, $\sum_{i=1}^{n}(x_i - \bar{x})^2$ est maximal lorsque $n/2$ valeurs de x valent x_{min} et $n/2$ valent x_{max}. Le plan optimal consiste à effectuer les essais par moitié[1] aux extrémités du domaine, ce qui contredit l'intuition de beaucoup de praticiens qui ont tendance à répartir régulièrement les valeurs de x dans l'intervalle de variation.

L'optimalité de ce plan est indissociable du modèle linéaire du premier degré. Si le modèle ne l'est pas et est par exemple du second degré $y = \beta_0 + \beta_1 x + \beta_2 (x)^2 + \varepsilon$, on ne pourra pas estimer β_2 : il est alors nécessaire d'introduire des essais au centre du domaine.

21.2.2 Plans orthogonaux pour *p* facteurs

Comme la régression linéaire multiple est invariante par changement d'échelle des variables, on notera -1 et 1 les valeurs minimales et maximales de chaque facteur (niveau bas et haut) ce qui revient à la transformation $\dfrac{x - \dfrac{(x_{min} + x_{max})}{2}}{\dfrac{x_{max} - x_{min}}{2}}$.

Sans contraintes sur le domaine, les expériences à réaliser se situeront aux sommets de l'hypercube, en raison de la propriété du paragraphe précédent et seuls les niveaux -1 et 1 seront utilisés.

La transformation en -1, 1 facilite grandement la vérification de l'orthogonalité de la matrice $\mathbf{X}$: $\mathbf{X'X} = n\mathbf{I}$. $\mathbf{X}$ doit être une matrice d'Hadamard qui n'existe que pour n multiple de 4.

Il faut tout d'abord que n soit pair : pour chaque facteur le nombre d'essais au niveau -1 doit être égal au nombre d'essais au niveau 1 pour avoir l'orthogonalité entre la colonne de 1 (associée au terme constant β_0) et la colonne associée à un facteur. Chaque colonne associée à un facteur est alors de moyenne nulle. Pour que deux colonnes associées à deux facteurs soient orthogonales, il faut que les 4 combinaisons de niveaux $(-1 ; -1)$ $(-1 ; 1)$ $(1 ; -1)$ $(1 ; 1)$ soient présentes le même nombre de fois : le plan est dit ***équilibré***.

21.2.2.1 Le plan factoriel complet

Il consiste à effectuer les 2^p expériences possibles. Il est orthogonal et donc D- et A- optimal. La matrice d'expériences pour le plan complet avec $p = 3$ est la suivante.

1 Si n est impair, on fait $(n-1)/2$ essais à chaque extrémité, et on met au hasard le $n$ième à une borne ou l'autre.

essai	A	B	C
1	−1	−1	−1
2	+1	−1	−1
3	−1	+1	−1
4	+1	+1	−1
5	−1	−1	+1
6	+1	−1	+1
7	−1	+1	+1
8	+1	+1	+1

La matrice du modèle **X** s'obtient en lui rajoutant la colonne de 1.

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & +1 & -1 & -1 \\ 1 & -1 & +1 & -1 \\ 1 & +1 & +1 & -1 \\ 1 & -1 & -1 & +1 \\ 1 & +1 & -1 & +1 \\ 1 & -1 & +1 & +1 \\ 1 & +1 & +1 & +1 \end{bmatrix}$$

(colonnes : I, A, B, C)

Remarquons que le modèle $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$ ne dépend que de 4 paramètres et que l'on a huit essais. Cela va permettre d'estimer sans essais supplémentaires des effets d'interaction A*B, A*C, B*C qui correspondent aux produits des variables :

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \beta_{12} x^1 x^2 + \beta_{13} x^1 x^3 + \beta_{23} x^2 x^3 + \varepsilon$$

La matrice associée à ce modèle s'obtient en rajoutant les colonnes obtenues en effectuant les produits terme à terme de deux colonnes parmi A, B, C. On vérifie que cette nouvelle matrice est encore orthogonale.

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 & +1 & +1 & +1 \\ 1 & +1 & -1 & -1 & -1 & -1 & +1 \\ 1 & -1 & +1 & -1 & -1 & +1 & -1 \\ 1 & +1 & +1 & -1 & +1 & -1 & -1 \\ 1 & -1 & -1 & +1 & +1 & -1 & -1 \\ 1 & +1 & -1 & +1 & -1 & +1 & -1 \\ 1 & -1 & +1 & +1 & -1 & -1 & +1 \\ 1 & +1 & +1 & +1 & +1 & +1 & +1 \end{bmatrix}$$

(colonnes : I, A, B, C, A*B, A*C, B*C)

On pourrait rajouter une 8ème colonne A*B*C mais le modèle est alors *saturé* car il y a autant d'essais que de paramètres à estimer et on ne pourra pas estimer la variance résiduelle. Notons à ce propos une confusion fréquente entretenue par les logiciels : ce qu'ils appellent variance résiduelle n'est autre que la somme des variances des effets considérés comme non significatifs. Pour véritablement estimer la variance résiduelle, il faut procéder à des essais supplémentaires (répétitions, points au centre).

21.2.2.2 Plans fractionnaires de type 2^{p-k} et plans de Plackett et Burman

Pour 4 facteurs, le plan complet demande 16 essais. Mais puisque la colonne A*B du plan précédent est orthogonale à toutes les autres, on peut l'attribuer à un quatrième facteur D. On aura alors un plan orthogonal, donc optimal, à 8 essais au lieu de 16 (demi-fraction).

A	B	C	D
−1	−1	−1	+1
+1	−1	−1	−1
−1	+1	−1	−1
+1	+1	−1	+1
−1	−1	+1	+1
+1	−1	+1	−1
−1	+1	+1	−1
+1	+1	+1	+1

On a perdu la possibilité d'estimer l'interaction A*B puisque celle ci est confondue avec le facteur D ; le plan est dit de *résolution* III. D'autres solutions sont possibles en attribuant D aux colonnes A*C ou B*C, la meilleure consiste à attribuer D à la colonne A*B*C car alors aucun effet principal n'est confondu avec une interaction entre deux facteurs mais seulement avec les interactions entre trois facteurs. On ne peut cependant estimer séparément les interactions d'ordre deux qui sont partiellement confondues entre elles ; le plan est de résolution IV. Le voici :

TABLEAU 21.1

essai	A	B	C	D
1	−1	−1	−1	−1
2	1	−1	−1	1
3	−1	1	−1	1
4	1	1	−1	−1
5	−1	−1	1	1
6	1	−1	1	−1
7	−1	1	1	−1
8	1	1	1	1

La structure de confusion des effets ou d'*alias* est :

1	A = BCD
2	B = ACD
3	C = ABD
4	D = ABC
5	AB = CD
6	AC = BD
7	AD = BC

Une autre façon de faire consiste à partir du plan en 32 essais et à prendre la moitié des essais, celle correspondant à A*B*C*D = 1. Le plan *complémentaire* ou *miroir* est obtenu en choisissant les autres essais A*B*C*D = −1 : il donne les expériences supplémentaires à faire si l'on veut estimer toutes les interactions (*désaliasage*) si l'analyse a montré que l'on ne pouvait pas négliger les interactions d'ordre 2.

Ces procédés de construction sont simples et bien connus ; ils aboutissent à des plans dont le nombre d'essais est une puissance de 2.

Le plan précédent en 8 essais peut convenir jusqu'à 7 facteurs (Tableau 21.2) en utilisant toutes les interactions, c'est un plan orthogonal à nombre d'essais minimal :

TABLEAU 21.2

Essai	A	B	C	D	E	F	G
1	−1	−1	−1	1	1	1	−1
2	1	−1	−1	−1	−1	1	1
3	−1	1	−1	−1	1	−1	1
4	1	1	−1	1	−1	−1	−1
5	−1	−1	1	1	−1	−1	1
6	1	−1	1	−1	1	−1	−1
7	−1	1	1	−1	−1	1	−1
8	1	1	1	1	1	1	1

Au delà de 8 facteurs on passe donc à 16 essais minimum. Les matrices d'Hadamard permettent de construire des plans orthogonaux, dits de Plackett et Burman dont le nombre d'essais est un multiple de 4 et est donc intermédiaire entre les puissances de 2. Pour 8 à 11 facteurs on pourra utiliser un plan en 12 essais tel celui donné par le tableau 21.3.

La structure de confusion des effets est très complexe (Montgomery, 2005) : chaque effet principal du plan précédent est partiellement confondu avec les 45 interactions d'ordre 2 ne le comprenant pas. Plus encore que les plans 2^{p-k}, ces plans doivent être utilisés avec précaution.

TABLEAU 21.3

Essai	A	B	C	D	E	F	G	H	I	J	K
1	−1	1	−1	1	1	1	−1	−1	−1	1	−1
2	−1	−1	1	−1	1	1	1	−1	−1	−1	1
3	1	−1	−1	1	−1	1	1	1	−1	−1	−1
4	−1	1	−1	−1	1	−1	1	1	1	−1	−1
5	−1	−1	1	−1	−1	1	−1	1	1	1	−1
6	−1	−1	−1	1	−1	−1	1	−1	1	1	1
7	1	−1	−1	−1	1	−1	−1	1	−1	1	1
8	1	1	−1	−1	−1	1	−1	−1	1	−1	1
9	1	1	1	−1	−1	−1	1	−1	−1	1	−1
10	−1	1	1	1	−1	−1	−1	1	−1	−1	1
11	1	−1	1	1	1	−1	−1	−1	1	−1	−1
12	1	1	1	1	1	1	1	1	1	1	1

Plan de Plackett et Burman en 12 essais pour 11 facteurs.

Les plans 2^{p-k} (dits de Box et Hunter) et de Plackett et Burman constituent des plans de ***criblage*** (« screening ») essentiellement destinés à éliminer rapidement des facteurs dans une étude préliminaire où de nombreux facteurs potentiels ont été soupçonnés.

21.2.3 Exemple

Un plan d'expériences a été réalisé selon la matrice du tableau 21.1 (données tirées de Montgomery 2001)

A	B	C	D	y
−1.0	−1.0	−1.0	−1.0	45
1.0	−1.0	−1.0	1.0	100
−1.0	1.0	−1.0	1.0	45
1.0	1.0	−1.0	−1.0	65
−1.0	−1.0	1.0	1.0	75
1.0	−1.0	1.0	−1.0	60
−1.0	1.0	1.0	−1.0	80
1.0	1.0	1.0	1.0	96

On calcule tout d'abord les effets des facteurs qui sont égaux aux différences des moyennes de la réponse entre le niveau +1 et le niveau −1 de chaque facteur ou interaction (tableau 21.4 et figure 21.1). Rappelons que AB est confondue avec CD, AC avec BD et AD avec BC.

On ne peut ici effectuer de test d'analyse de la variance car il n'y a pas assez de degrés de liberté pour estimer la variance résiduelle.

Il est clair que le facteur B est sans effet ainsi que les interactions AB et CD qui peuvent être éliminés. On négligera également BD et BC qui ne peuvent être séparées de AC et AD en faisant l'hypothèse que si le facteur B n'a pas d'effet principal, on peut ne pas tenir compte des interactions entre B et les autres facteurs.

TABLEAU 21.4

Effets estimés pour réponse

moyenne	=	70.75
A : Facteur_A	=	19.0
B : Facteur_B	=	1.5
C : Facteur_C	=	14.0
D : Facteur_D	=	16.5
AB + CD	=	−1.0
AC + BD	=	−18.5
AD + BC	=	19.0

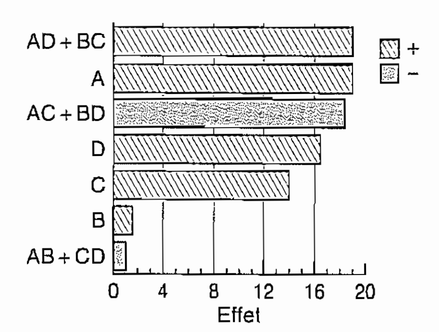


FIGURE 21.1

TABLEAU 21.5

Analyse de la variance pour y

Source	Somme des carrés	DDL	Carré moyen	F	Proba.
A : Facteur_A	722.0	1	722.0		
B : Facteur_B	4.5	1	4.5		
C : Facteur_C	392.0	1	392.0		
D : Facteur_D	544.5	1	544.5		
AB + CD	2.0	1	2.0		
AC + BD	684.5	1	684.5		
AD + BC	722.0	1	722.0		
Erreur totale	0.0	0			
Total (corr.)	3071.5	7			

On réestime alors le modèle simplifié, écrit symboliquement $Y = I + A + C + D + AC + AD$. Tous les effets sont significatifs. L'orthogonalité laisse invariantes les sommes de carrés. L'« erreur totale » est en fait la somme des carrés négligés.

TABLEAU 21.6

Analyse de la variance pour y

Source	Somme des carrés	DDL	carré moyen	F	Proba.
A :Facteur_A	722.0	1	722.0	222.15	0.0045
C :Facteur_C	392.0	1	392.0	120.62	0.0082
D :Facteur_D	544.5	1	544.5	167.54	0.0059
AC	684.5	1	684.5	210.62	0.0047
AD	722.0	1	722.0	222.15	0.0045
Erreur totale	6.5	2	3.25		
Total (corr.)	3071.5	7			

Le modèle de régression final s'écrit :

y = 70.75 + 9.5 A + 7.0C + 8.25D−9.25AC + 9.5AD

21.3 QUELQUES PLANS POUR SURFACES DE RÉPONSE DU SECOND DEGRÉ

Il s'agit de trouver des matrices d'essais pour des modèles linéaires avec des termes de degré 2 comme celui-ci : $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 (x^1)^2 + \beta_4 (x^2)^2 + \beta_5 x^1 x^2 + \varepsilon$.

Il faut donner à chaque facteur au moins trois niveaux pour pouvoir estimer les effets du second degré. Lorsque le domaine expérimental est cubique, ces trois niveaux seront définis par les extrêmes et le milieu de l'intervalle de variation de chaque facteur et recodés en −1, 0, 1 après la transformation déjà vue plus haut $\dfrac{x - \dfrac{(x_{\min} + x_{\max})}{2}}{\dfrac{x_{\max} - x_{\min}}{2}}$.

Il ne peut exister de plans orthogonaux pour de tels modèles et la recherche s'est focalisée sur des plans possédant d'autres propriétés comme l'isovariance par rotation. La possibilité d'expérimentation séquentielle est également très utile ; elle consiste à augmenter un plan factoriel fractionnaire de criblage permettant d'estimer des effets principaux en lui ajoutant des points au centre et d'autres points pour estimer les autres effets.

Il existe bien d'autres plans que ceux présentés maintenant parmi les plus classiques, et nous renvoyons aux ouvrages déjà cités. L'analyse des résultats d'expérience se fait avec la régression linéaire multiple.

21.3.1 Plans composites à faces centrées

Ce nom s'explique de la manière suivante. Pour 3 facteurs le domaine expérimental est un cube. On effectue tout d'abord les 8 essais aux sommets du cube, que l'on complète par 6 essais aux centres des faces, et n_c essais au centre du cube.

Voici la matrice d'expériences pour $n_c = 2$:

TABLEAU 21.7

essai	A	B	C
1	−1.0	−1.0	−1.0
2	1.0	−1.0	−1.0
3	−1.0	1.0	−1.0
4	1.0	1.0	−1.0
5	−1.0	−1.0	1.0
6	1.0	−1.0	1.0
7	−1.0	1.0	1.0
8	1.0	1.0	1.0
9	−1.0	0.0	0.0
10	1.0	0.0	0.0
11	0.0	−1.0	0.0
12	0.0	1.0	0.0
13	0.0	0.0	−1.0
14	0.0	0.0	1.0
15	0.0	0.0	0.0
16	0.0	0.0	0.0

Plan composite à faces centrées pour 3 facteurs

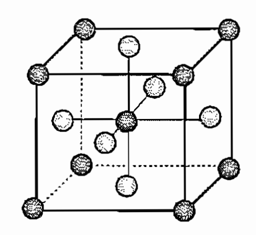


FIGURE 21.2

Ce plan n'est pas orthogonal (il ne peut pas l'être) : voici la matrice de corrélation entre les colonnes de **X** :

	A	B	C	A²	AB	AC	B²	BC	C²
A	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
C	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
A²	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.4667	0.0000	0.4667
AB	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
AC	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
B²	0.0000	0.0000	0.0000	0.4667	0.0000	0.0000	1.0000	0.0000	0.4667
BC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
C²	0.0000	0.0000	0.0000	0.4667	0.0000	0.0000	0.4667	0.0000	1.0000

Il n'est pas non plus isovariant par rotation. Ces plans se généralisent à un nombre quelconque de facteurs. L'hypercube a 2^p sommets et $2p$ faces. La partie factorielle peut-être une fraction orthogonale et non le plan complet. Le plan minimal pour 5 facteurs comprendra en tout 28 essais avec 2 points au centre, en partant d'un plan 2^{5-1} et laissera 7 degrés de liberté.

21.3.2 Plans composites généraux

Au lieu de mettre les points en « étoile » au centre des faces, ils sont à une distance α du centre. Ce qui donne le plan suivant pour 3 facteurs et 2 points au centre. Il y a donc 5 niveaux par facteur.

TABLEAU 21.8

essai	A	B	C
1	−1.0	−1.0	−1.0
2	1.0	−1.0	−1.0
3	−1.0	1.0	−1.0
4	1.0	1.0	−1.0
5	−1.0	−1.0	1.0
6	1.0	−1.0	1.0
7	−1.0	1.0	1.0
8	1.0	1.0	1.0
9	$-\alpha$	0.0	0.0
10	α	0.0	0.0
11	0.0	$-\alpha$	0.0
12	0.0	α	0.0
13	0.0	0.0	$-\alpha$
14	0.0	0.0	α
15	0.0	0.0	0.0
16	0.0	0.0	0.0

Plan composite centré pour 3 facteurs

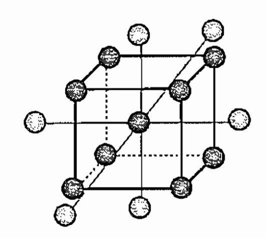


FIGURE 21.3

On montre que pour obtenir l'isovariance, il faut prendre $\alpha = (n_f)^{1/4}$ oú n_f est le nombre d'essais de la partie factorielle. Pour 3 facteurs $\alpha = 8^{0.25} = 1.6818$ et pour 2 facteurs $\alpha = 4^{0.25} = 2^{0.5} = 1.414$. Les points sont alors respectivement sur une sphère ou un cercle. Ces plans conviennent donc bien quand le domaine expérimental est sphérique.

21.3.3 Plans de Box-Behnken

Ce sont des plans où les facteurs ne prennent que les niveaux -1, 0, 1. Pour $p = 3$ les essais hors du centre sont disposés au milieu des arêtes du cube (figure 21.4), pour $p > 3$ au milieu des hyperfaces de dimension p-1 . Ces plans demandent souvent moins d'essais que les composites. Tous les points (hors ceux au centre) sont situés sur une sphère de carré de rayon égal à 2 si p = 3, 4 ou 5, de carré de rayon égal à 3 pour p = 6 ou 7. Ils ne contiennent aucun sommet ce qui peut-être un intérêt si les sommets correspondent à des expériences difficiles à réaliser.

Le plus utilisé est celui pour 3 facteurs donné par le tableau 21.9 . Il n'est pas isovariant par rotation. La matrice de corrélation entre effets (tableau 21.10) montre une nette supériorité sur le plan composite à faces centrées étudié plus haut.

TABLEAU 21.9

essai	A	B	C
1	−1.0	−1.0	0.0
2	1.0	−1.0	0.0
3	−1.0	1.0	0.0
4	1.0	1.0	0.0
5	−1.0	0.0	−1.0
6	1.0	0.0	−1.0
7	−1.0	0.0	1.0
8	1.0	0.0	1.0
9	0.0	−1.0	−1.0
10	0.0	1.0	−1.0
11	0.0	−1.0	1.0
12	0.0	1.0	1.0
13	0.0	0.0	0.0
14	0.0	0.0	0.0
15	0.0	0.0	0.0

Plan de Box-Behnken pour 3 facteurs

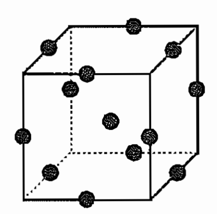


FIGURE 21.4

TABLEAU 21.10

	A	B	C	A^2	AB	AC	B^2	BC	C^2
A	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
C	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
A^2	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	−0.0714	0.0000	−0.0714
AB	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
AC	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
B^2	0.0000	0.0000	0.0000	−0.0714	0.0000	0.0000	1.0000	0.0000	−0.0714
BC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
C^2	0.0000	0.0000	0.0000	−0.0714	0.0000	0.0000	−0.0714	0.0000	1.0000

Matrice des corrélations entre effets du Plan de Box-Behnken

Le tableau 21.11 est un plan de Box-Behnken pour 4 facteurs en 27 essais dont 3 au centre. Ce plan est isovariant par rotation.

TABLEAU 21.11

essai	A	B	C	D
1	−1.0	−1.0	0.0	0.0
2	1.0	−1.0	0.0	0.0
3	−1.0	1.0	0.0	0.0
4	1.0	1.0	0.0	0.0
5	0.0	0.0	−1.0	−1.0
6	0.0	0.0	1.0	−1.0
7	0.0	0.0	1.0	1.0
8	0.0	0.0	1.0	1.0
9	−1.0	0.0	0.0	−1.0
10	1.0	0.0	0.0	−1.0
11	−1.0	0.0	0.0	1.0
12	1.0	0.0	0.0	1.0
13	0.0	−1.0	−1.0	0.0
14	0.0	1.0	−1.0	0.0
15	0.0	−1.0	1.0	0.0
16	0.0	1.0	1.0	0.0
17	−1.0	0.0	−1.0	0.0
18	1.0	0.0	−1.0	0.0
19	−1.0	0.0	1.0	0.0
20	1.0	0.0	1.0	0.0
21	0.0	−1.0	0.0	−1.0
22	0.0	1.0	0.0	−1.0
23	0.0	−1.0	0.0	1.0
24	0.0	1.0	0.0	1.0
25	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0

Plan de Box-Behnken pour 4 facteurs

21.3.4 Application à un problème d'optimisation

Les données sont reprises de Montgomery (2001 page 503) avec le plan de Box-Behnken du tableau 21.9.

Y' = (535 580 596 563 645 458 350 600 595 648 532 656 653 599 620)

Analyse de la variance pour y

Source	Somme des carrés	DDL	Carré moyen.	F	Proba.
A :Facteur_A	703.125	1	703.125	0.67	0.4491
B :Facteur_B	6105.13	1	6105.13	5.85	0.0602
C :Facteur_C	5408.0	1	5408.0	5.18	0.0719
AA	20769.2	1	20769.2	19.90	0.0066
AB	1521.0	1	1521.0	1.46	0.2814
AC	47742.3	1	47742.3	45.74	0.0011
BB	1404.0	1	1404.0	1.35	0.2985
BC	1260.25	1	1260.25	1.21	0.3219
CC	4719.0	1	4719.0	4.52	0.0868
Erreur totale	5218.75	5	1043.75		
Total (corr.)	94871.3	14			

L'analyse de la variance et le graphe des effets indiquent que l'on peut éliminer le terme du premier degré en A, celui du deuxième degré en B ainsi que les produits AB et BC.

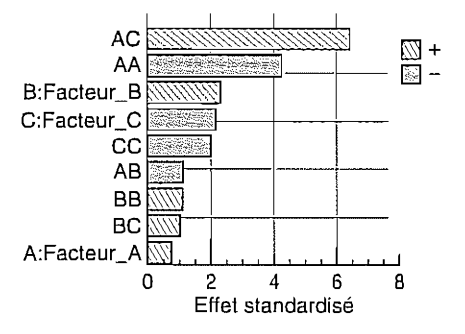


FIGURE 21.5

Le modèle restreint est significatif et s'écrit :

$$y = 636.0 + 27.625B - 26.0C - 76.5A^2 + 109.25AC - 37.25C^2$$

Analyse de la variance pour y

Source	Somme des carrés	DDL	Carré moyen.	F	Proba.
B :Facteur_B	6105.13	1	6105.13	5.44	0.0446
C :Facteur_C	5408.0	1	5408.0	4.82	0.0558
AA	21736.9	1	21736.9	19.36	0.0017
AC	47742.3	1	47742.3	42.51	0.0001
CC	5153.8	1	5153.8	4.59	0.0608
Erreur totale	10107.1	9	1123.01		
Total (corr.)	94871.3	14			

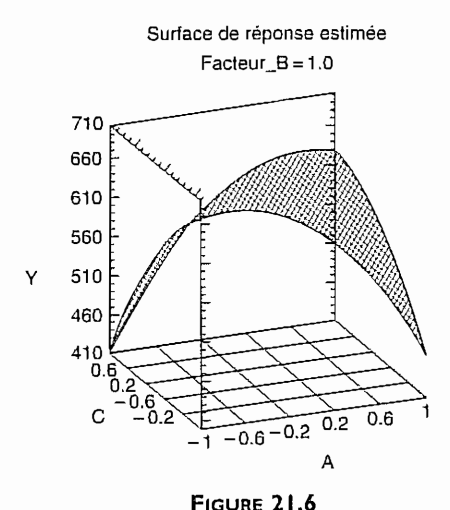


FIGURE 21.6

La figure 21.6 représente pour B fixé au niveau 1 la surface de réponse de y selon A et C. Si l'on cherche à maximiser la réponse, compte tenu des contraintes du domaine l'optimum est atteint pour $A = -0.71$, $B = 1$ et $C = -1$ et vaut environ 691.

21.4 PLANS POUR FACTEURS QUALITATIFS

Notons m_j le nombre de modalités du facteur n°j. Bien que non ordonnées ces modalités seront encore appelées « niveaux ». Leur numérotation est donc arbitraire et les niveaux seront notés tantôt par les entiers $1, 2, \ldots, m_j$, tantôt par des symboles $A1, A2, \ldots A_{mj}$.

21.4.1 Orthogonalités

Comme précédemment, l'orthogonalité est une propriété souvent recherchée. L'orthogonalité d'un plan pour un modèle donné se traduit par une analyse de variance orthogonale : les sommes de carrés des différents effets sont additives.

Une condition suffisante d'orthogonalité pour le modèle à effets principaux sans intéraction est que le plan soit *équilibré* au sens suivant : pour toute paire de facteurs i et j les $m_i m_j$ traitements sont présents le même nombre de fois. On parle également d'orthogonalité au ***sens strict*** ; elle entraîne la D-optimalité. Cela implique que le nombre d'essais soit un multiple de $m_i m_j$. Toutes les cases du tableau croisé à m_i lignes et m_j colonnes de dénombrement des essais ont le même effectif.

Il y a orthogonalité au ***sens large*** si les effectifs ne sont pas identiques (plan non équilibré) mais si le khi-deux calculé sur ce tableau est nul. L'analyse de la variance a les mêmes propriétés que pour l'orthogonalité stricte mais le plan n'est pas nécessairement D-optimal.

21.4.2 Facteurs à *m* niveaux

Si tous les facteurs ont le même nombre de niveaux m, le plan complet nécessite m^p expériences. Nous nous intéresserons ici aux plans nécessitant moins d'observations.

Si $m = 2$, on peut utiliser les mêmes plans que dans le cas quantitatif : factoriels fractionnaires, Plackett et Burman. La seule différence est que les niveaux « bas » et « haut » n'ont pas de sens et que la modélisation ne s'exprimera pas à l'aide d'une régression linéaire classique mais plutôt symboliquement comme suit :

$$y = \beta_0 + \begin{pmatrix} \beta_1 \\ -\beta_1 \end{pmatrix} + \begin{pmatrix} \beta_2 \\ -\beta_2 \end{pmatrix} + \ \dots \ + \begin{pmatrix} \beta_p \\ -\beta_p \end{pmatrix} + \varepsilon$$

On ajoute β_j si le facteur j est au niveau 1 et $-\beta_j$ s'il est au niveau 2.

Pour 3 et 4 facteurs on utilisera les carrés latins et gréco-latins, au delà on se reportera à des recueils de table (par exemple celles de Benoist & al. 1994), ou on les construira par des procédés algorithmiques

21.4.2.1 Carrés latins

Pour $p = 3$ et m quelconque, les plans en carrés latins sont des plans orthogonaux au sens strict en m^2 essais au lieu de m^3. On peut les obtenir de la manière suivante, d'où leur nom :

On constitue un carré en croisant 2 des 3 facteurs, et on affecte à chaque case les niveaux du 3ème facteur par permutations circulaires de la première ligne. Chaque niveau de chaque facteur est associé une fois et une seule à chaque niveau d'un des deux autres. Voici le carré latin pour 3 facteurs à 4 niveaux qui comprend 16 essais noté parfois $L_{16}4^3$. Le premier essai est A1 B1 C1 etc.

	B1	**B2**	**B3**	**B4**
A1	C1	C2	C3	C4
A2	C2	C3	C4	C1
A3	C3	C4	C1	C2
A4	C4	C1	C2	C3

La matrice des essais de ce plan en notations classiques est :

essai	A	B	C
1	1	1	1
2	1	2	2
3	1	3	3
4	1	4	4
5	2	1	2
6	2	2	3
7	2	3	4
8	2	4	1
9	3	1	3
10	3	2	4
11	3	3	1
12	3	4	2
13	4	1	4
14	4	2	1
15	4	3	2
16	4	4	3

Les carrés latins ne peuvent estimer que les effets principaux. Dans l'analyse de variance le nombre de degrés de liberté pour l'erreur vaut $m^2 - 1 - 3(m - 1) = (m - 1)(m - 2)$. Il y a en effet $(m - 1)$ paramètres à estimer pour chaque facteur plus le terme constant.

21.4.2.2 *Carrés gréco-latins*

Ce sont des plans pour 4 facteurs à m niveaux. On peut les construire en superposant deux carrés latins. Ils existent pour tout $m > 2$ sauf pour $m = 6$. Les plans en carrés gréco-latins sont des plans orthogonaux au sens strict en m^2 essais au lieu de m^4.

Voici le carré gréco-latin $L_{16}4^4$ sous sa forme originelle et sous forme de matrice d'expérience :

	B1	**B2**	**B3**	**B4**
A1	C1 α	C2 β	C3 γ	C4 δ
A2	C2 γ	C1 δ	C4 α	C3 β
A3	C3 δ	C4 γ	C1 β	C2 α
A4	C4 β	C3 α	C2 δ	C1 γ

Comme pour les carrés latins, on ne peut estimer que les effets principaux ; dans l'analyse de variance le nombre de degrés de liberté pour l'erreur vaut $(m - 1)(m - 3)$.

essai	A	B	C	D
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	1	4	4	4
5	2	1	3	3
6	2	2	4	4
7	2	3	1	1
8	2	4	2	2
9	3	1	4	4
10	3	2	3	3
11	3	3	2	2
12	3	4	1	1
13	4	1	2	2
14	4	2	1	1
15	4	3	4	4
16	4	4	3	3

21.4.3 Plans asymètriques

On désigne ainsi les plans avec des facteurs n'ayant pas tous le même nombre de niveaux. La construction de plans fractionnaires est difficile et il n'existe pas de méthode générale. Voici tout d'abord quelques indications concernant le nombre d'essais :

Il doit être au minimum égal au nombre de paramètres à estimer $n > \sum_{j=1}^{p} (m_j - 1) + 1$ et pour avoir un plan orthogonal au sens strict (sans interaction) être un multiple commun de tous les produits $m_i m_j$. Cela donne quelques possibilités, mais il n'est pas certain qu'un tel plan existe en dehors du plan complet.

On peut alors consulter des recueils de tables, mais ils ne sont pas exhaustifs, ou utiliser une des méthodes suivantes : fusion ou compression en partant d'autres plans.

En voici deux exemples simples :

21.4.3.1 Un exemple de fusion

On cherche un plan pour 3 facteurs, l'un à 4 niveaux, les deux autres à 2 niveaux. Le plan complet comporte 16 essais. Le ppcm des produits $m_i m_j$ vaut 8. On trouve le plan de la manière suivante : on part de 4 facteurs à deux niveaux et du plan fractionnaire 2^{4-1} du tableau 21.1 :

essai	A	B	C	D
1	−1	−1	−1	−1
2	1	−1	−1	1
3	−1	1	−1	1
4	1	1	−1	−1
5	−1	−1	1	1
6	1	−1	1	−1
7	−1	1	1	−1
8	1	1	1	1

En combinant deux facteurs à 2 niveaux on en obtient un à 4 niveaux. On remplace par exemple les colonnes C et D de la façon suivante (−1 ; −1) devient le niveau 1 d'un facteur E, (−1, 1) le niveau 2, (1 ; −1) le niveau 3 et (1 ; 1) le niveau 4. Le plan résultant est strictement orthogonal et permet d'estimer les effets principaux.

essai	A	B	E
1	−1	−1	1
2	1	−1	2
3	−1	1	2
4	1	1	1
5	−1	−1	4
6	1	−1	3
7	−1	1	3
8	1	1	4

21.4.3.2 *Un exemple de compression*

Cette technique consiste à regrouper des niveaux d'un facteur (« collapsing » en anglais).

Cherchons un plan pour 3 facteurs : A et B à 3 niveaux et C à deux niveaux. Le plan complet demande 12 essais, mais seuls les effets principaux sont à estimer. Le ppcm de 6 et 4 est également 12, donc le plan complet est le seul plan orthogonal au sens strict. Si le facteur C avait eu 3 niveaux, on aurait pu utiliser un carré latin 3^3 comme celui-ci :

	B1	**B2**	**B3**
A1	C1	C2	C3
A2	C2	C3	C1
A3	C3	C1	C2

Il suffit alors de regrouper deux des 3 niveaux de C ; par exemple C3 et C2. On obtient un plan orthogonal, mais cette fois-ci au sens large puisque non équilibré.

	B1	**B2**	**B3**
A1	C1	C2	C2
A2	C2	C2	C1
A3	C2	C1	C2

essai	A	B	C
1	1	1	1
2	1	2	2
3	1	3	2
4	2	1	2
5	2	2	2
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	2

21.5 CONSTRUCTION ALGORITHMIQUE DE PLANS OPTIMAUX

Il n'est pas toujours possible de construire « à la main » un plan d'expériences, soit parce que le domaine expérimental est irrégulier, ou qu'il est impossible de trouver un plan orthogonal par les méthodes exposées précédemment (il n'y en a pas ou on ne sait pas le trouver), ou encore parce que le nombre d'essais est limité.

On utlisera alors des logiciels, maintenant assez répandus, pour trouver un plan optimal, en général D-optimal. On se donne un ensemble de N points candidats ou essais potentiels parcourant le domaine expérimental, parmi lesquels on va chercher un sous-ensemble de n essais (avec répétitions ou non). En l'absence de contraintes sur le domaine, N correspond au nombre d'essais du plan complet. Comme il est irréalisable d'explorer tous les choix de n parmi N, on utilise des algorithmes d'optimisation basés pour les plus connus sur des échanges : on part d'un plan, souvent choisi aléatoirement, que l'on améliore en échangeant un essai du plan contre un autre non choisi. Ces algorithmes ne convergent pas nécessairement vers l'optimum du critère et il est conseillé de les relancer plusieurs fois en faisant varier les initialisations.

S'il existe un plan orthogonal pour la valeur fixée de n et si l'algorithme converge, alors il découvrira ce plan. Sinon on trouvera un plan de bonne qualité.

Exemple : On a 4 facteurs A, B, C, D à 3, 4, 2, 2 niveaux respectivement. On ne s'intéresse qu'aux effets principaux. Le plan complet possède 48 essais. Il existe un plan orthogonal au sens strict en 24 essais mais il est trop onéreux et on veut se contenter de 12 essais ce qui laissera encore 4 degrés de liberté pour l'erreur résiduelle. En 10 itérations on aboutit au plan suivant :

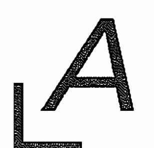
Essai	A	B	C	D
1	3	4	2	2
2	3	3	2	1
3	3	2	1	2
4	3	1	1	1
5	2	4	2	1
6	2	3	1	1
7	2	2	1	2
8	2	1	2	2
9	1	4	1	2
10	1	3	2	2
11	1	2	2	1
12	1	1	1	1

Il y a orthogonalité entre A et B, A et C, A et D, C et D mais pas entre B et C, ni entre B et D.

L'efficacité d'un plan D-optimal est souvent mesurée par la quantité $\frac{|\mathbf{X'X}|^{1/p}}{n}$ que l'on interprète comme le rapport entre le nombre hypothètique d'essais d'un plan orthogonal qui aurait même déterminant et le nombre d'essais du plan. En effet on sait que pour un plan orthogonal $\mathbf{X'X}$ est diagonale de termes tous égaux à l'effectif du plan. La D-efficacité vaut ici 97.0984 %.

On peut également « forcer » certains essais (que l'on veut faire, ou qui ont déjà été réalisés) et optimiser sur les essais restant à faire. Les algorithmes de plans D-optimaux constituent une solution pratique, mais ne sont pas une panacée : la solution optimale peut être instable, certains essais trop complexes, l'optimum du déterminant ne correspond pas forcément au critère recherché . . .

Annexes



Tables usuelles

Table A.1 : Nombres au hasard.
Table A.2 : Loi binomiale : fonction de répartition.
Table A.3 : Loi binomiale : probabilités individuelles.
Table A.3 bis : Abaque pour les intervalles de confiance d'une proportion p.
Table A.4 : Loi de Poisson.
Table A.5 : Loi normale centrée-réduite : fonction de répartition.
Table A.5 bis : Loi normale centrée-réduite : inverse de la fonction de répartition.
Table A.6 : Loi du khi-deux.
Table A.7 : Loi de Fisher-Snedecor.
Table A.8 : Loi de Student.
Table A.9 : Valeurs critiques du coefficient de corrélation.
Table A.9 bis : Abaque pour les intervalles de confiance d'un coefficient de corrélation.
Table A.10 : Corrélation transformée de Fisher.
Table A.11 : Valeurs critiques du coefficient de corrélation des rangs de Spearman.
Table A.12 : Test de concordance de p classements (W de Kendall).
Table A.13 : Loi de la statistique de Cramer-von Mises.
Table A.14 : Valeurs critiques pour le test de Kolmogorov.
Table A.15 : Valeurs critiques du coefficient d'asymétrie.
Table A.16 : Valeurs critiques du coefficient d'aplatissement.
Table A.17 : Test de Durbin et Watson.
Table A.18 : Coefficients pour calculer l'espérance et la variance de l'écart-type corrigé et de l'étendue d'un échantillon gaussien.

Les tables A.1 et A.2 sont extraites de J. Mothes, *Prévisions et décisions statistiques dans l'entreprise*, Dunod, 1968.

Les tables A.3 et A.9 *bis* sont extraites de Massey et Dixon, *Introduction to statistical analysis*, Mc Graw-Hill, 1951.

L'abaque A.3 *bis* est extrait de E. Morice et F. Chartier, *Méthode statistique*, deuxième partie, *INSEE*, 1954.

Les tables A.6 et A.7 sont extraites de Hald, *Statistical tables and formulas*, Wiley, 1952.

La table A.9 est extraite des tables scientifiques éditées par Ciba-Geigy, 1973.

La table A.11 est extraite d'un article de J. H. Zar paru dans le *Journal of the American Statistical Association*, n° 339 de septembre 1972.

La table A.12 est adaptée de celle de M. G. Kendall, *Rank correlation methods*, Ch. Griffin and Co., 1962.

La table A.13 est extraite d'un article de Knott paru dans *Journal of the Royal Statistical Society*, B36, n° 3, p. 436, 1974.

La table A.14 est extraite d'un article de L. H. Miller paru dans *Journal of the American Statistical Association*, 51, pp. 113–115, 1956.

Les tables A.4, A.5, A.5 *bis*, A.8, A.10 sont extraites du numéro spécial de la *Revue de Statistique Appliquée*, édite par l'Institut de Statistique des Universités de Paris, 1973.

Les tables A.15 et A.16 sont extraites de E. S. Pearson et H. O. Hartley *Biometrika tables for statisticians*, 2 tomes, Cambridge University Press, 1969–1972, qui contient de nombreuses autres tables spécialisées.

La table A.17 est extraite de Chatterjee-Price *Regression Analysis by Example*, Wiley, New York, 1977.

L'index bibliographique de Greenwood et Hartley, *Guide of tables in mathematical statistics*, 1014 pages, Princeton University Press, 1962, est une précieuse liste de références.

TABLE A.1 NOMBRES AU HASARD

	5	10	15	20	25	30	35	40	45	50
	13407	62899	78937	90525	25033	56358	78902	47008	72488	57949
	50230	63237	94083	93634	71652	02656	57532	60307	91619	48916
	84980	62458	09703	78397	66179	46982	67619	39254	90763	74056
	22116	33646	17545	31321	65772	86506	09811	82848	92211	51178
5	68645	15068	56898	87021	40115	27524	42221	88293	67592	06430
	26518	39122	96561	56004	50260	68648	85596	83979	09041	62350
	36493	41666	27871	71329	69212	57932	65281	57233	07732	58439
	77402	12994	59892	85581	70823	53338	34405	67080	16568	00854
	83679	97154	40341	84741	08967	73287	94952	59008	95774	44927
10	71802	39356	02981	89107	79788	51330	37129	31898	34011	43304
	57494	72484	22676	44311	15356	05348	03582	66183	68392	86844
	73364	38416	93128	10297	11419	82937	84389	88273	96010	09843
	14499	83965	75403	18002	45068	54257	18085	92625	60911	39137
	40747	03084	07734	88940	88722	85717	73810	79866	84853	68647
15	42237	59122	92855	62097	81276	06318	81607	00565	56626	77422
	32934	60227	58707	44858	36081	79981	01291	68707	45427	82145
	05764	14284	73069	80830	17231	42936	48472	18782	51646	37564
	32706	94879	93188	66049	25988	46656	35365	13800	83745	40141
	22190	27559	95668	53261	21676	98943	43618	42110	93402	93997
20	81616	15641	94921	95970	63506	22007	29966	38144	62556	07864
	26099	65801	69870	84446	58248	21282	56938	54729	67757	68412
	71874	61692	80001	21430	02305	59741	34262	15157	27545	14522
	08774	29689	42245	51903	69179	96682	91819	60812	47631	50609
	37294	92028	56850	83380	05912	29830	37612	15593	73198	99287
25	33912	37996	78967	57201	66916	73998	54289	07147	84313	51938
	63610	61475	26980	23804	54972	72068	19403	53756	04281	98022
	01570	41701	30282	54647	06077	29354	95704	75928	21811	88274
	24159	77787	38973	82178	46802	90245	01805	23906	96559	06785
	92834	52941	88301	22127	23459	40229	74678	21859	98645	72388
30	16178	60063	59284	16279	48003	44634	08623	32752	40742	05470
	81808	32980	80660	98391	62243	19678	39551	18398	36918	43543
	28628	82072	04854	52809	86608	68017	11120	28638	72850	03650
	62249	65757	12273	91261	96983	15082	83851	77682	81728	52157
	84541	99891	01585	96711	29712	02877	70955	59693	26838	96011
35	89052	39061	99811	69831	47234	93263	47386	17462	18874	74210

TABLE A.2 LOI BINOMIALE

Fonction de répartition $P_k = \sum_0^k C_n^k p^k (1-p)^{n-k}$

Taille de l'échantillon	k	$p = 1\ \%$	$p = 2\ \%$	$p = 3\ \%$	$p = 4\ \%$	$p = 5\ \%$	$p = 6\ \%$	$p = 7\ \%$	$p = 8\ \%$	$p = 9\ \%$	$p = 10\ \%$	$p = 20\ \%$	$p = 30\ \%$	$p = 40\ \%$	$p = 50\ \%$
	0	0,9510	0,9039	0,8587	0,8153	0,7738	0,7339	0,6957	0,6591	0,6240	0,5905	0,3277	0,1681	0,0778	0,0313
	1	0,9980	0,9962	0,9915	0,9852	0,9774	0,9681	0,9575	0,9466	0,9326	0,9185	0,7373	0,5282	0,3370	0,1875
$N = 5$	2	1	1	0,9997	0,9994	0,9988	0,9980	0,9969	0,9955	0,9937	0,9914	0,9421	0,8369	0,6826	0,5000
	3			1	1	1	1	0,9999	0,9998	0,9997	0,9995	0,9933	0,9692	0,9130	0,8125
	4							1	1	1	1	0,9997	0,9976	0,9898	0,9687
	5											1	1	1	1
	0	0,9044	0,8171	0,7374	0,6648	0,5987	0,5386	0,4840	0,4344	0,3894	0,3487	0,1074	0,0282	0,0060	0,0010
	1	0,9957	0,9838	0,9655	0,9418	0,9139	0,8824	0,8483	0,8121	0,7746	0,7361	0,3758	0,1493	0,0464	0,0107
	2	0,9999	0,9991	0,9972	0,9938	0,9885	0,9812	0,9717	0,9599	0,9460	0,9298	0,6778	0,3828	0,1673	0,0547
	3	1	1	0,9999	0,9996	0,9990	0,9980	0,9964	0,9942	0,9912	0,9872	0,8791	0,6496	0,3823	0,1719
	4			1	1	0,9999	0,9998	0,9997	0,9994	0,9990	0,9984	0,9672	0,8497	0,6331	0,3770
$N = 10$	5					1	1	1	1	0,9999	0,9999	0,9936	0,9527	0,8338	0,6230
	6									1	1	0,9991	0,9894	0,9452	0,8281
	7											0,9999	0,9984	0,9877	0,9453
	8											1	0,9999	0,9983	0,9893
	9												1	0,9999	0,9990
	10													1	1
	0	0,8601	0,7386	0,6333	0,5421	0,4633	0,3953	0,3367	0,2863	0,2430	0,2059	0,0352	0,0047	0,0005	0,0000
	1	0,9904	0,9647	0,9270	0,8809	0,8290	0,7738	0,7168	0,6597	0,6035	0,5490	0,1671	0,0353	0,0052	0,0005
	2	0,9996	0,9970	0,9906	0,9797	0,9638	0,9429	0,9171	0,8870	0,8531	0,8159	0,3980	0,1268	0,0271	0,0037
	3	1	0,9998	0,9992	0,9976	0,9945	0,9896	0,9825	0,9727	0,9601	0,9445	0,6482	0,2969	0,0905	0,0176
	4		1	0,9999	0,9998	0,9994	0,9986	0,9972	0,9950	0,9918	0,9873	0,8358	0,5155	0,2173	0,0592
	5			1	1	1	0,9999	0,9997	0,9993	0,9987	0,9978	0,9389	0,7216	0,4032	0,1509
	6						1	1	0,9999	0,9999	0,9997	0,9819	0,8689	0,6098	0,3036
	7								1	1	1	0,9958	0,9500	0,7869	0,5000
$N = 15$	8											0,9992	0,9848	0,9050	0,6964
	9											0,9999	0,9963	0,9662	0,8491
	10											1	0,9993	0,9907	0,9408
	11												0,9999	0,9981	0,9824
	12												1	0,9997	0,9963
	13													1	0,9995
	14														1
	15														

TABLE A.2 (*suite*) LOI BINOMIALE

Fonction de répartition $P_k = \sum_{0}^{k} C_n^k p^k (1 - p)^{n-k}$

Taille de l'échantillon	k	$p = 1$ %	$p = 2$ %	$p = 3$ %	$p = 4$ %	$p = 5$ %	$p = 6$ %	$p = 7$ %	$p = 8$ %	$p = 9$ %	$p = 10$ %	$p = 20$ %	$p = 30$ %	$p = 40$ %	$p = 50$ %
	0	0,8179	0,6676	0,5438	0,4420	0,3585	0,2901	0,2342	0,1887	0,1516	0,1216	0,0115	0,0008	—	—
	1	0,9831	0,9401	0,8802	0,8103	0,7358	0,6605	0,5869	0,5169	0,4516	0,3917	0,0692	0,0076	0,0005	—
	2	0,9990	0,9929	0,9790	0,9561	0,9245	0,8850	0,8390	0,7879	0,7334	0,6769	0,2061	0,0355	0,0036	0,0002
	3	1	0,9994	0,9973	0,9926	0,9841	0,9710	0,9529	0,9294	0,9007	0,8670	0,4114	0,1071	0,0160	0,0013
	4		1	0,9997	0,9990	0,9974	0,9944	0,9893	0,9817	0,9710	0,9568	0,6296	0,2375	0,0510	0,0059
	5			1	0,9999	0,9997	0,9991	0,9981	0,9962	0,9932	0,9887	0,8042	0,4164	0,1256	0,0207
	6				1	1	0,9999	0,9997	0,9994	0,9987	0,9976	0,9133	0,6080	0,2500	0,0577
	7						1	1	0,9999	0,9998	0,9996	0,9679	0,7723	0,4159	0,1316
	8								1	1	0,9999	0,9900	0,8867	0,5956	0,2517
	9										1	0,9974	0,9520	0,7553	0,4119
$N = 20$	10											0,9994	0,9829	0,8725	0,5881
	11											0,9999	0,9949	0,9435	0,7483
	12											1	0,9987	0,9790	0,8684
	13												0,9997	0,9935	0,9423
	14												1	0,9984	0,9793
	15													0,9997	0,9941
	16													1	0,9987
	17														0,9998
	18														1
	19														
	20														

Table A.2 (*suite*) Loi binomiale

Fonction de répartition $P_k = \sum_0^k C_n^k p^k (1-p)^{n-k}$

Taille de l'échantillon	k	$p = 1$ %	$p = 2$ %	$p = 3$ %	$p = 4$ %	$p = 5$ %	$p = 6$ %	$p = 7$ %	$p = 8$ %	$p = 9$ %	$p = 10$ %	$p = 20$ %	$p = 30$ %	$p = 40$ %	$p = 50$ %
	0	0,7397	0,5455	0,4010	0,2939	0,2146	0,1563	0,1134	0,0820	0,0591	0,0424	0,0012	0,0000	—	—
	1	0,9639	0,8794	0,7731	0,6612	0,5535	0,4555	0,3694	0,2958	0,2343	0,1837	0,0105	0,0003	—	—
	2	0,9967	0,9783	0,9399	0,8831	0,8122	0,7324	0,6488	0,5654	0,4855	0,4114	0,0442	0,0021	0,0000	—
	3	0,9998	0,9971	0,9881	0,9694	0,9392	0,8974	0,8450	0,7842	0,7175	0,6474	0,1227	0,0093	0,0003	—
	4	0,9999	0,9996	0,9982	0,9937	0,9844	0,9685	0,9447	0,9126	0,8723	0,8245	0,2552	0,0302	0,0015	0,0000
	5	1	1	0,9997	0,9989	0,9967	0,9921	0,9838	0,9707	0,9519	0,9268	0,4275	0,0766	0,0057	0,0002
	6			1	0,9999	0,9994	0,9983	0,9960	0,9918	0,9848	0,9742	0,6070	0,1595	0,0172	0,0007
	7				1	0,9999	0,9997	0,9992	0,9980	0,9959	0,9922	0,7608	0,2814	0,0435	0,0026
	8					1	0,9999	0,9999	0,9996	0,9990	0,9980	0,8713	0,4315	0,0940	0,0081
	9						1	1	0,9999	0,9998	0,9995	0,9389	0,5888	0,1763	0,0214
	10								1	1	0,9999	0,9744	0,7304	0,2915	0,0494
	11										1	0,9905	0,8407	0,4311	0,1002
	12											0,9969	0,9155	0,5785	0,1808
	13											0,9991	0,9599	0,7145	0,2923
	14											0,9998	0,9831	0,8246	0,4278
$N = 30$	15											1	0,9936	0,9029	0,5722
	16												0,9979	0,9519	0,7077
	17												0,9994	0,9798	0,8192
	18												0,9998	0,9917	0,8998
	19												1	0,9971	0,9506
	20													0,9991	0,9786
	21													0,9998	0,9919
	22													1	0,9974
	23														0,9993
	24														0,9998
	25														1
	26														
	27														
	28														
	29														
	30														

Table A.2 (*suite*) Loi binomiale

Fonction de répartition $P_k = \sum_0^k C_n^k p^k (1-p)^{n-k}$

Taille de l'échantillon	k	$p = 1$ %	$p = 2$ %	$p = 3$ %	$p = 4$ %	$p = 5$ %	$p = 6$ %	$p = 7$ %	$p = 8$ %	$p = 9$ %	$p = 10$ %	$p = 20$ %	$p = 30$ %	$p = 40$ %	$p = 50$ %
	0	0,6690	0,4457	0,2957	0,1954	0,1285	0,0842	0,0549	0,0356	0,0230	0,0148	0,0001	—	—	—
	1	0,9393	0,8095	0,6615	0,5210	0,3991	0,2990	0,2201	0,1594	0,1140	0,0805	0,0015	—	—	—
	2	0,9925	0,9543	0,8822	0,7855	0,6767	0,5665	0,4625	0,3694	0,2894	0,2228	0,0079	0,0001	—	—
	3	0,9993	0,9918	0,9686	0,9252	0,8619	0,7827	0,3837	0,6007	0,5092	0,4231	0,0285	0,0006	—	—
	4	1	0,9988	0,9933	0,9790	0,9520	0,9104	0,8546	0,7868	0,7103	0,6290	0,0759	0,0026	—	—
	5		0,9999	0,9988	0,9951	0,9861	0,9691	0,9419	0,9033	0,8535	0,7937	0,1613	0,0086	0,0001	—
	6		1	0,9998	0,9990	0,9966	0,9909	0,9801	0,9624	0,9361	0,9005	0,2859	0,0238	0,0006	—
	7			1	0,9998	0,9993	0,9977	0,9942	0,9873	0,9758	0,9581	0,4371	0,0553	0,0021	—
	8				1	0,9999	0,9995	0,9985	0,9963	0,9920	0,9845	0,5931	0,1100	0,0061	0,0001
	9					1	0,9999	0,9997	0,9990	0,9976	0,9949	0,7318	0,1959	0,0156	0,0003
	10						1	0,9999	0,9998	0,9994	0,9985	0,8392	0,3087	0,0352	0,0011
	11							1	1	0,9999	0,9996	0,9125	0,4406	0,0709	0,0032
	12									1	0,9999	0,9568	0,5772	0,1285	0,0083
	13										1	0,9806	0,7032	0,2112	0,0192
	14											0,9921	0,8074	0,3174	0,0403
	15											0,9971	0,8849	0,4402	0,0769
$N = 40$	16											0,9990	0,9367	0,5681	0,1341
	17											0,9997	0,9680	0,6885	0,2148
	18											0,9999	0,9852	0,7911	0,3179
	19											1	0,9937	0,8702	0,4373
	20												0,9976	0,9256	0,5627
	21												0,9991	0,9608	0,6821
	22												0,9997	0,9811	0,7852
	23												0,9999	0,9917	0,8659
	24												1	0,9966	0,9231
	25													0,9988	0,9597
	26													0,9996	0,9808
	27													0,9999	0,9917
	28													1	0,9968
	29														0,9989
	30														0,9997
	31														0,9999
	32														1
	33														
	34														

Fonction de répartition $P_k = \sum_{0}^{k} C_n^k p^k (1-p)^{n-k}$

Taille de l'échantillon	k	$p = 1$ %	$p = 2$ %	$p = 3$ %	$p = 4$ %	$p = 5$ %	$p = 6$ %	$p = 7$ %	$p = 8$ %	$p = 9$ %	$p = 10$ %	$p = 20$ %	$p = 30$ %	$p = 40$ %	$p = 50$ %
	0	0,6050	0,3642	0,2181	0,1299	0,0769	0,0453	0,0266	0,0155	0,0090	0,0052	—	—	—	—
	1	0,9106	0,7358	0,5553	0,4005	0,2794	0,1900	0,1265	0,0827	0,0532	0,0338	0,0002	—	—	—
	2	0,9862	0,9216	0,8108	0,6767	0,5405	0,4162	0,3108	0,2260	0,1605	0,1117	0,0013	—	—	—
	3	0,9984	0,9822	0,9372	0,8609	0,7604	0,6473	0,5327	0,4253	0,3303	0,2503	0,0057	0,0000	—	—
	4	0,9999	0,9968	0,9832	0,9510	0,8964	0,8206	0,7290	0,6290	0,5277	0,4312	0,0185	0,0002	—	—
	5	1	0,9995	0,9963	0,9856	0,9622	0,9224	0,8650	0,7919	0,7072	0,6161	0,0480	0,0007	—	—
	6		0,9999	0,9993	0,9964	0,9882	0,9711	0,9417	0,8981	0,8404	0,7702	0,1034	0,0025	0,0000	—
	7		1	0,9999	0,9992	0,9968	0,9906	0,9780	0,9562	0,9232	0,8779	0,1904	0,0073	0,0001	—
	8			1	0,9999	0,9992	0,9973	0,9927	0,9834	0,9672	0,9421	0,3073	0,0183	0,0002	—
	9				1	0,9998	0,9993	0,9978	0,9944	0,9875	0,9755	0,4437	0,0402	0,0008	—
	10					1	0,9998	0,9994	0,9983	0,9957	0,9906	0,5836	0,0809	0,0022	—
	11						1	0,9999	0,9995	0,9987	0,9968	0,7107	0,1390	0,0057	0,0000
	12							1	0,9999	0,9996	0,9990	0,8139	0,2229	0,0133	0,0002
	13								1	0,9999	0,9997	0,8894	0,3279	0,0280	0,0005
	14									1	0,9999	0,9393	0,4468	0,0540	0,0013
	15										1	0,9692	0,5692	0,0955	0,0033
	16											0,9856	0,6839	0,1561	0,0077
	17											0,9937	0,7822	0,2369	0,0164
	18											0,9975	0,8594	0,3356	0,0325
$N = 50$	19											0,9991	0,9152	0,4465	0,0595
	20											0,9997	0,9522	0,5610	0,1013
	21											0,9999	0,9749	0,6701	0,1611
	22											1	0,9877	0,7660	0,2399
	23												0,9944	0,8438	0,3359
	24												0,9976	0,9022	0,4439
	25												0,9991	0,9427	0,5561
	26												0,9997	0,9686	0,6641
	27												0,9999	0,9840	0,7601
	28												1	0,9924	0,8389
	29													0,9966	0,8987
	30													0,9986	0,9405
	31													0,9995	0,9675
	32													0,9998	0,9836
	33													0,9999	0,9923
	34													1	0,9967
	35														0,9987
	36														0,9995
	37														0,9998
	38														1

TABLE A.3 PROBABILITÉS BINOMIALES $C_N^x p^x (1-p)^{N-X}$ POUR $N \leq 10$ ET POUR DIVERSES VALEURS DE p

N	X \ p	.01	.05	.10	.15	.20	.25	.30	$\frac{1}{3}$	.35	.40	.45	.50
2	0	.9801	.9025	.8100	.7225	.6400	.5625	.4900	.4444	.4225	.3600	.3025	.2500
	1	.0198	.0950	.1800	.2550	.3200	.3750	.4200	.4444	.4550	.4800	.4950	.5000
	2	.0001	.0025	.0100	.0225	.0400	.0625	.0900	.1111	.1225	.1600	.2025	.2500
3	0	.9703	.8574	.7290	.6141	.5120	.4219	.3430	.2963	.2746	.2160	.1664	.1250
	1	.0294	.1354	.2430	.3251	.3840	.4219	.4410	.4444	.4436	.4320	.4084	.3750
	2	.0003	.0071	.0270	.0574	.0960	.1406	.1890	.2222	.2389	.2880	.3341	.3750
	3	.0000	.0001	.0010	.0034	.0080	.0156	.0270	.0370	.0429	.0640	.0911	.1250
4	0	.9606	.8145	.6561	.5220	.4096	.3164	.2401	.1975	.1785	.1296	.0915	.0625
	1	.0388	.1715	.2916	.3685	.4096	.4219	.4116	.3951	.3845	.3456	.2995	.2500
	2	.0006	.0135	.0486	.0975	.1536	.2109	.2646	.2963	.3105	.3456	.3675	.3750
	3	.0000	.0005	.0036	.0115	.0256	.0469	.0756	.0988	.1115	.1536	.2005	.2500
	4	.0000	.0000	.0001	.0005	.0016	.0039	.0081	.0123	.0150	.0256	.0410	.0625
5	0	.9510	.7738	.5905	.4437	.3277	.2373	.1681	.1317	.1160	.0778	.0503	.0312
	1	.0480	.2036	.3280	.3915	.4096	.3955	.3602	.3292	.3124	.2592	.2059	.1562
	2	.0010	.0214	.0729	.1382	.2048	.2637	.3087	.3292	.3364	.3456	.3369	.3125
	3	.0000	.0011	.0081	.0244	.0512	.0879	.1323	.1646	.1811	.2304	.2757	.3125
	4	.0000	.0000	.0004	.0022	.0064	.0146	.0284	.0412	.0488	.0768	.1128	.1562
	5	.0000	.0000	.0000	.0001	.0003	.0010	.0024	.0041	.0053	.0102	.0185	.0312
6	0	.9415	.7351	.5314	.3771	.2621	.1780	.1176	.0878	.0754	.0467	.0277	.0156
	1	.0571	.2321	.3543	.3993	.3932	.3560	.3025	.2634	.2437	.1866	.1359	.0938
	2	.0014	.0305	.0984	.1762	.2458	.2966	.3241	.3292	.3280	.3110	.2780	.2344
	3	.0000	.0021	.0146	.0415	.0819	.1318	.1852	.2195	.2355	.2765	.3032	.3125
	4	.0000	.0001	.0012	.0055	.0154	.0330	.0595	.0823	.0951	.1382	.1861	.2344
	5	.0000	.0000	.0001	.0004	.0015	.0044	.0102	.0165	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0014	.0018	.0041	.0083	.0156
7	0	.9321	.6983	.4783	.3206	.2097	.1335	.0824	.0585	.0490	.0280	.0152	.0078
	1	.0659	.2573	.3720	.3960	.3670	.3115	.2471	.2048	.1848	.1306	.0872	.0547
	2	.0020	.0406	.1240	.2097	.2753	.3115	.3177	.3073	.2985	.2613	.2140	.1641
	3	.0000	.0036	.0230	.0617	.1147	.1730	.2269	.2561	.2679	.2903	.2918	.2734
	4	.0000	.0002	.0026	.0109	.0287	.0577	.0972	.1280	.1442	.1935	.2388	.2734
	5	.0000	.0000	.0002	.0012	.0043	.0115	.0250	.0384	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0000	.0001	.0004	.0013	.0036	.0064	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0006	.0016	.0037	.0078
8	0	.9227	.6634	.4305	.2725	.1678	.1001	.0576	.0390	.0319	.0168	.0084	.0039
	1	.0746	.2793	.3826	.3847	.3355	.2670	.1977	.1561	.1373	.0896	.0548	.0312
	2	.0026	.0515	.1488	.2376	.2936	.3115	.2965	.2731	.2587	.2090	.1569	.1094
	3	.0001	.0054	.0331	.0839	.1468	.2076	.2541	.2731	.2786	.2787	.2568	.2188
	4	.0000	.0004	.0046	.0185	.0459	.0865	.1361	.1707	.1875	.2322	.2627	.2734
	5	.0000	.0000	.0004	.0026	.0092	.0231	.0467	.0683	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0000	.0002	.0011	.0038	.0100	.0171	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0000	.0001	.0004	.0012	.0024	.0033	.0079	.0164	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0002	.0007	.0017	.0039
9	0	.9135	.6302	.3874	.2316	.1342	.0751	.0404	.0260	.0207	.0101	.0046	.0020
	1	.0830	.2985	.3874	.3679	.3020	.2253	.1556	.1171	.1004	.0605	.0339	.0176
	2	.0034	.0629	.1722	.2597	.3020	.3003	.2668	.2341	.2162	.1612	.1110	.0703
	3	.0001	.0077	.0446	.1069	.1762	.2336	.2668	.2731	.2716	.2508	.2119	.1641
	4	.0000	.0006	.0074	.0283	.0661	.1168	.1715	.2048	.2194	.2508	.2600	.2461
	5	.0000	.0000	.0008	.0050	.0165	.0389	.0735	.1024	.1181	.1672	.2128	.2461
	6	.0000	.0000	.0001	.0006	.0028	.0087	.0210	.0341	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0000	.0003	.0012	.0039	.0073	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0009	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003	.0008	.0020
10	0	.9044	.5987	.3487	.1969	.1074	.0563	.0282	.0173	.0135	.0060	.0025	.0010
	1	.0914	.3151	.3874	.3474	.2684	.1877	.1211	.0867	.0725	.0403	.0207	.0098
	2	.0042	.0746	.1937	.2759	.3020	.2816	.2335	.1951	.1757	.1209	.0763	.0439
	3	.0001	.0105	.0574	.1298	.2013	.2503	.2668	.2601	.2522	.2150	.1665	.1172
	4	.0000	.0010	.0112	.0401	.0881	.1460	.2001	.2276	.2377	.2508	.2384	.2051
	5	.0000	.0001	.0015	.0085	.0264	.0584	.1029	.1366	.1536	.2007	.2340	.2461
	6	.0000	.0000	.0001	.0012	.0055	.0162	.0368	.0569	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0000	.0001	.0008	.0031	.0090	.0163	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0014	.0030	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010

TABLE A.3 *bis* ABAQUE DONNANT EN FONCTION DE f L'INTERVALLE DE CONFIANCE À 0.95 ($p_{0.025}$ À $p_{0.975}$)

f : fréquence observée (en %) sur un échantillon d'effectif n

p : proportion (en %) dans la population échantillonnée

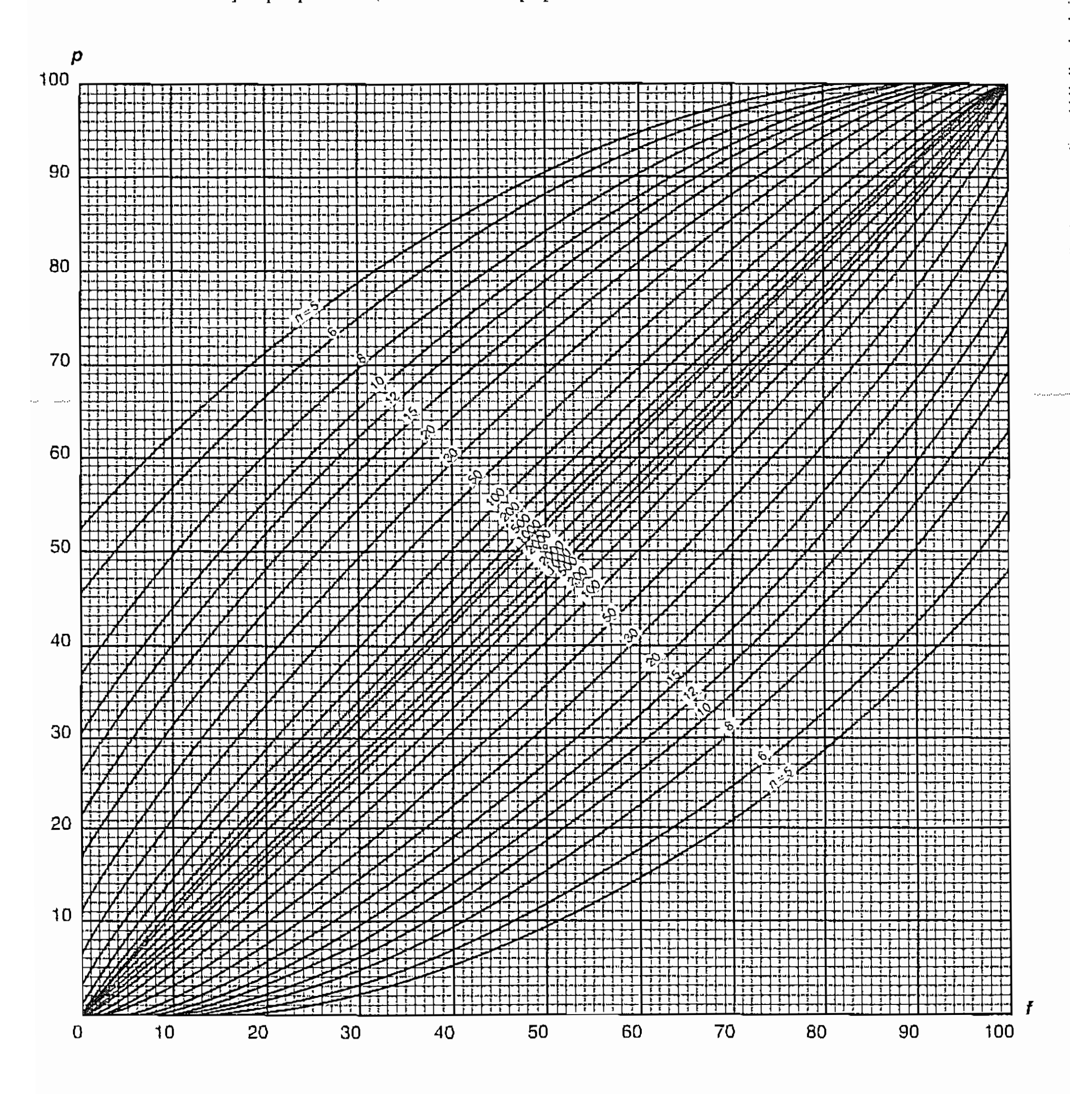


TABLE A.4 LOI DE POISSON

	Probabilités individuelles $P(X = k) = e^{-m} \frac{m^k}{k!}$								
k	$m = 0,1$	$m = 0,2$	$m = 0,3$	$m = 0,4$	$m = 0,5$	$m = 0,6$	$m = 0,7$	$m = 0,8$	$m = 0,9$
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494
4		0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111
5				0,0001	0,0002	0,0004	0,0007	0,0012	0,0020
6							0,0001	0,0002	0,0003

	Probabilités cumulées $P(X \leq c) = \sum_{k=0}^{k=c} e^{-m} \frac{m^k}{k!}$								
c	$m = 0,1$	$m = 0,2$	$m = 0,3$	$m = 0,4$	$m = 0,5$	$m = 0,6$	$m = 0,7$	$m = 0,8$	$m = 0,9$
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,9953	0,9825	0,9631	0,9384	0,9098	0,8781	0,8442	0,8088	0,7725
2	0,9998	0,9988	0,9964	0,9920	0,9856	0,9769	0,9659	0,9526	0,9372
3	1	0,9999	0,9997	0,9992	0,9982	0,9966	0,9942	0,9909	0,9866
4		1	1	0,9999	0,9998	0,9996	0,9992	0,9986	0,9977
5				1	1	1	0,9999	0,9998	0,9997
6							1	1	1

Remarques :

1) Si X suit une loi de Poisson de paramètre m on a la relation exacte :

$$P(X \leq c) = P(\chi^2_{2(c+1)} > 2m)$$

2) Si m est > 18 on peut utiliser l'approximation grossière :

$$\frac{X + 0.5 - m}{\sqrt{m}} \simeq U$$

où U est la variable de Laplace-Gauss centrée-réduite.

3) Une approximation plus précise est donnée par :

$$P(X \leq c) \simeq P\left(U > 3\sqrt{c+1}\left(\left(\frac{m}{c+1}\right)^{1/3} + \frac{1}{9(c+1)} - 1\right)\right)$$

TABLE A.4 (*suite*) LOI DE POISSON

k	Probabilités individuelles $P(X = k) = e^{-m}\frac{m^k}{k!}$								
	$m = 1,0$	$m = 1,5$	$m = 2,0$	$m = 2,5$	$m = 3,0$	$m = 3,5$	$m = 4,0$	$m = 4,5$	$m = 5,0$
0	0,3679	0,2231	0,1353	0,0821	0,0498	0,0302	0,0183	0,0111	0,0067
1	0,3679	0,3347	0,2707	0,2052	0,1494	0,1057	0,0733	0,0500	0,0337
2	0,1839	0,2510	0,2707	0,2565	0,2240	0,1850	0,1465	0,1125	0,0842
3	0,0613	0,1255	0,1804	0,2138	0,2240	0,2158	0,1954	0,1687	0,1404
4	0,0153	0,0471	0,0902	0,1336	0,1680	0,1888	0,1954	0,1898	0,1755
5	0,0031	0,0141	0,0361	0,0668	0,1008	0,1322	0,1563	0,1708	0,1755
6	0,0005	0,0035	0,0120	0,0278	0,0504	0,0771	0,1042	0,1281	0,1462
7	0,0001	0,0008	0,0034	0,0099	0,0216	0,0385	0,0595	0,0824	0,1044
8		0,0001	0,0009	0,0031	0,0081	0,0169	0,0298	0,0463	0,0653
9			0,0002	0,0009	0,0027	0,0066	0,0132	0,0232	0,0363
10				0,0002	0,0008	0,0023	0,0053	0,0104	0,0181
11					0,0002	0,0007	0,0019	0,0043	0,0082
12					0,0001	0,0002	0,0006	0,0016	0,0034
13						0,0001	0,0002	0,0006	0,0013
14							0,0001	0,0002	0,0005
15								0,0001	0,0002
16									0,0001

c	Probabilités cumulées $P(X \leq c) = \sum_{k=0}^{k=c} e^{-m}\frac{m^k}{k!}$								
	$m = 1,0$	$m = 1,5$	$m = 2,0$	$m = 2,5$	$m = 3,0$	$m = 3,5$	$m = 4,0$	$m = 4,5$	$m = 5,0$
0	0,3679	0,2231	0,1353	0,0821	0,0498	0,0302	0,0183	0,0111	0,0067
1	0,7358	0,5578	0,4060	0,2873	0,1991	0,1359	0,0916	0,0611	0,0404
2	0,9197	0,8088	0,6767	0,5438	0,4232	0,3208	0,2381	0,1736	0,1247
3	0,9810	0,9344	0,8571	0,7576	0,6472	0,5366	0,4335	0,3423	0,2650
4	0,9963	0,9814	0,9473	0,8912	0,8153	0,7254	0,6288	0,5321	0,4405
5	0,9994	0,9955	0,9834	0,9579	0,9161	0,8576	0,7851	0,7029	0,6160
6	0,9999	0,9991	0,9955	0,9858	0,9665	0,9347	0,8893	0,8311	0,7622
7	1	0,9998	0,9989	0,9958	0,9881	0,9733	0,9489	0,9134	0,8666
8		1	0,9998	0,9989	0,9962	0,9901	0,9786	0,9597	0,9319
9			1	0,9997	0,9989	0,9967	0,9919	0,9829	0,9682
10				0,9999	0,9997	0,9990	0,9972	0,9933	0,9863
11				1	0,9999	0,9997	0,9991	0,9976	0,9945
12					1	0,9999	0,9997	0,9992	0,9980
13						1	0,9999	0,9997	0,9993
14							1	0,9999	0,9998
15								1	0,9999
16									1

TABLE A.4 (*suite*) LOI DE POISSON

k	Probabilités individuelles $P(X = k) = e^{-m}\frac{m^k}{k!}$								
	m = 5,5	m = 6,0	m = 6.5	m = 7.0	m = 7,5	m = 8,0	m = 8,5	m = 9,0	m = 9,5
0	0,0041	0.0025	0,0015	0,0009	0,0006	0,0003	0.0002	0,0001	0,0001
1	0,0225	0,0149	0,0098	0,0064	0.0041	0,0027	0.0017	0,0011	0.0007
2	0.0618	0,0446	0,0318	0,0223	0,0156	0.0107	0,0074	0.0050	0,0034
3	0,1133	0,0892	0.0688	0,0521	0,0389	0,0286	0,0208	0,0150	0.0107
4	0.1558	0,1339	0.1118	0,0912	0,0729	0,0573	0,0443	0,0337	0,0254
5	0,1714	0,1606	0.1454	0,1277	0,1094	0,0916	0,0752	0,0607	0,0483
6	0,1571	0.1606	0,1575	0.1490	0,1367	0,1221	0.1066	0.0911	0,0764
7	0,1234	0.1377	0,1462	0,1490	0,1465	0.1396	0.1294	0,1171	0.1037
8	0.0849	0,1033	0,1188	0,1304	0,1373	0,1396	0,1375	0,1318	0,1232
9	0.0519	0,0688	0,0858	0,1014	0,1144	0.1241	0,1299	0,1318	0,1300
10	0,0285	0,0413	0,0558	0,0710	0,0858	0,0993	0,1104	0,1186	0,1235
11	0.0143	0,0225	0,0330	0,0452	0,0585	0,0722	0,0853	0,0970	0.1067
12	0,0065	0,0113	0,0179	0,0264	0,0366	0,0481	0,0604	0,0728	0,0844
13	0.0028	0,0052	0,0089	0.0142	0.0211	0,0296	0.0395	0,0504	0,0617
14	0,0011	0,0022	0,0041	0,0071	0,0113	0,0169	0,0240	0,0324	0,0419
15	0.0004	0,0009	0.0018	0,0033	0.0057	0,0090	0.0136	0,0194	0,0265
16	0,0001	0,0003	0,0007	0,0014	0,0026	0,0045	0,0072	0,0109	0,0157
17		0,0001	0,0003	0,0006	0,0012	0,0021	0,0036	0,0058	0,0088
18			0,0001	0.0002	0.0005	0,0009	0,0017	0,0029	0,0046
19				0,0001	0,0002	0,0004	0,0008	0,0014	0.0023
20					0,0001	0,0002	0,0003	0.0006	0,0011
21						0.0001	0,0001	0,0003	0,0005
22							0,0001	0,0001	0,0002
23									0,0001
24									

TABLE A.4 (*suite*) LOI DE POISSON

c	Probabilités cumulées $P(X \leq c) = \sum_{k=0}^{k=c} e^{-m} \frac{m^k}{k!}$								
	$m = 5,5$	$m = 6,0$	$m = 6,5$	$m = 7,0$	$m = 7,5$	$m = 8,0$	$m = 8,5$	$m = 9,0$	$m = 9,5$
0	0.0041	0,0025	0,0015	0,0009	0,0006	0,0003	0,0002	0.0001	0,0001
1	0.0266	0,0174	0,0113	0.0073	0,0047	0.0030	0,0019	0,0012	0,0008
2	0,0884	0.0620	0,0430	0,0296	0.0203	0,0138	0,0093	0,0062	0,0042
3	0,2017	0,1512	0.1118	0,0818	0,0591	0,0424	0,0301	0,0212	0,0149
4	0,3575	0,2851	0.2237	0,1730	0.1321	0.0996	0,0746	0.0550	0,0403
5	0,5289	0,4457	0,3690	0,3007	0,2414	0,1912	0,1496	0,1157	0,0885
6	0,6860	0.6063	0.5265	0,4497	0,3782	0,3134	0,2562	0,2068	0,1649
7	0,8095	0,7440	0,6728	0,5987	0,5246	0.4530	0,3856	0.3239	0.2687
8	0.9044	0,8472	0,7916	0,7291	0.6620	0,5925	0,5231	0,4557	0.3918
9	0,9462	0,9161	0,8774	0,8305	0,7764	0,7166	0,6530	0,5874	0,5218
10	0.9747	0.9574	0.9332	0,9015	0,8622	0,8159	0,7634	0,7060	0,6453
11	0,9890	0,9799	0,9661	0,9466	0,9208	0,8881	0,8487	0,8030	0,7520
12	0,9955	0,9912	0.9840	0,9730	0,9573	0,9362	0,9091	0,8758	0,8364
13	0,9983	0,9964	0,9929	0,9872	0,9784	0.9658	0.9486	0,9261	0.8981
14	0,9994	0,9986	0,9970	0.9943	0,9897	0,9827	0,9726	0,9585	0,9400
15	0,9998	0,9995	0,9988	0.9976	0,9954	0,9918	0,9862	0,9780	0.9665
16	0.9999	0.9998	0,9996	0,9990	0,9980	0,9963	0,9934	0,9889	0,9823
17	1	1	0,9998	0,9996	0,9992	0,9984	0,9970	0,9947	0,9911
18			1	0,9999	0,9997	0,9993	9,9987	0,9976	0.9957
19				1	0,9999	0.9997	0.9995	0,9989	0,9980
20					1	0,9999	0,9998	0,9996	0,9991
21						1	0.9999	0.9998	0,9996
22							1	0,9999	0,9998
23								1	0.9999
24									1

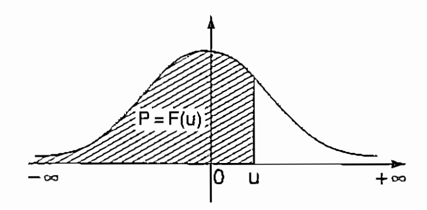
TABLE A.4 (*suite*) LOI DE POISSON

k	Probabilités individuelles $P(X = k) = e^{-m} \frac{m^k}{k!}$								
	$m = 10$	$m = 11$	$m = 12$	$m = 13$	$m = 14$	$m = 15$	$m = 16$	$m = 17$	$m = 18$
0									
1	0,0005	0.0002	0.0001						
2	0,0023	0.0010	0.0004	0.0002	0,0001				
3	0.0076	0.0037	0.0018	0.0008	0,0004	0.0002	0.0001		
4	0,0189	0.0102	0.0053	0.0027	0.0013	0.0007	0,0003	0.0002	0.0001
5	0,0378	0,0224	0,0127	0,0070	0,0037	0,0019	0.0010	0.0005	0.0002
6	0.0631	0,0411	0.0255	0,0152	0.0087	0.0048	0.0026	0.0014	0.0007
7	0.0901	0.0646	0,0437	0.0281	0,0174	0.0104	0,0060	0,0034	0.0019
8	0,1126	0,0888	0.0655	0,0457	0.0304	0.0194	0,0120	0.0072	0,0042
9	0,1251	0.1085	0,0874	0.0661	0,0473	0.0324	0,0213	0,0135	0.0083
10	0,1251	0,1194	0.1048	0,0859	0.0663	0,0486	0.0341	0,0230	0.0150
11	0.1137	0.1194	0,1144	0.1015	0.0844	0,0663	0.0496	0,0356	0,0245
12	0,0948	0,1094	0,1144	0,1099	0,0984	0.0829	0.0661	0.0504	0.0368
13	0.0729	0,0926	0.1056	0,1099	0,1060	0.0956	0,0814	0,0658	0.0509
14	0,0521	0,0728	0,0905	0,1021	0.1060	0,1024	0.0930	0.0700	0,0655
15	0,0347	0,0534	0,0724	0.0885	0,0989	0,1024	0.0992	0,0906	0.0786
16	0.0217	0,0367	0.0543	0,0719	0,0866	0,0960	0.0992	0,0963	0.0884
17	0.0128	0.0237	0,0383	0.0550	0,0713	0,0847	0.0934	0,0963	0,0936
18	0,0071	0,0145	0.0255	0.0397	0.0554	0,0706	0,0830	0.0909	0,0936
19	0.0037	0.0084	0,0161	0,0272	0,0409	0.0558	0.0699	0,0814	0,0887
20	0,0019	0,0046	0.0097	0.0177	0.0286	0,0418	0.0559	0.0692	0,0798
21	0,0009	0.0024	0.0055	0.0109	0,0191	0.0299	0.0426	0.0560	0,0684
22	0,0004	0,0012	0.0030	0.0065	0.0121	0,0204	0.0310	0.0433	0,0560
23	0,0002	0,0006	0.0016	0.0037	0,0074	0.0133	0.0216	0,0320	0.0438
24	0,0001	0.0003	0.0008	0.0020	0.0043	0,0083	0.0144	0.0227	0,0329
25		0,0001	0.0004	0.0010	0.0024	0,0050	0.0092	0.0154	0.0237
26			0.0002	0.0005	0.0013	0,0029	0,0057	0.0101	0,0164
27			0.0001	0.0002	0.0007	0.0016	0.0034	0,0063	0.0109
28				0.0001	0.0003	0,0009	0,0019	0.0039	0.0070
29					0.0002	0.0004	0.0010	0,0023	0.0044
30					0.0001	0,0002	0.0006	0.0013	0.0026
31						0,0001	0.0003	0.0007	0.0015
32						0.0001	0,0001	0.0004	0.0009
33							0.0001	0.0002	0.0005
34								0.0001	0.0002
35									0.0001
36									0.0001

TABLE A.4 (*suite et fin*) LOI DE POISSON

c	Probabilités cumulées $P(X \leq c) = \sum_{k=0}^{k=c} e^{-m} \frac{m^k}{k!}$								
	$m = 10$	$m = 11$	$m = 12$	$m = 13$	$m = 14$	$m = 15$	$m = 16$	$m = 17$	$m = 18$
0									
1	0,0005	0,0002	0,0001						
2	0,0028	0,0012	0,0005	0,0002	0,0001				
3	0,0104	0,0049	0,0023	0,0010	0,0005	0,0002	0,0001		
4	0,0293	0,0151	0,0076	0,0037	0,0018	0,0009	0,0004	0,0002	0,0001
5	0,0671	0,0375	0,0203	0,0107	0,0055	0,0028	0,0014	0,0007	0,0003
6	0,1302	0,0786	0,0458	0,0259	0,0142	0,0076	0,0040	0,0021	0,0010
7	0,2203	0,1432	0,0895	0,0540	0,0316	0,0180	0,0100	0,0054	0,0029
8	0,3329	0,2320	0,1550	0,0997	0,0620	0,0374	0,0220	0,0126	0,0071
9	0,4580	0,3405	0,2424	0,1658	0,1093	0,0698	0,0433	0,0261	0,0154
10	0,5831	0,4599	0,3472	0,2517	0,1756	0,1184	0,0774	0,0491	0,0304
11	0,6968	0,5793	0,4616	0,3532	0,2600	0,1847	0,1270	0,0847	0,0549
12	0,7916	0,6887	0,5760	0,4631	0,3584	0,2676	0,1931	0,1350	0,0917
13	0,8645	0,7813	0,6816	0,5730	0,4644	0,3622	0,2745	0,2009	0,1426
14	0,9166	0,8541	0,7721	0,6751	0,5704	0,4656	0,3675	0,2808	0,2081
15	0,9513	0,9075	0,8445	0,7636	0,6693	0,5680	0,4667	0,3714	0,2867
16	0,9730	0,9442	0,8988	0,8355	0,7559	0,6640	0,5659	0,4677	0,3750
17	0,9857	0,9679	0,9371	0,8905	0,8272	0,7487	0,6593	0,5440	0,4686
18	0,9928	0,9824	0,9626	0,9302	0,8826	0,8193	0,7423	0,6550	0,5622
19	0,9965	0,9908	0,9787	0,9574	0,9235	0,8751	0,8122	0,7363	0,6509
20	0,9984	0,9954	0,9884	0,9751	0,9521	0,9169	0,8681	0,8055	0,7307
21	0,9993	0,9978	0,9939	0,9860	0,9712	0,9468	0,9107	0,8615	0,7991
22	0,9997	0,9990	0,9969	0,9925	0,9833	0,9672	0,9617	0,9048	0,8551
23	0,9999	0,9996	0,9985	0,9962	0,9907	0,9805	0,9633	0,9367	0,8989
24	1	0,9999	0,9993	0,9982	0,9950	0,9888	0,9777	0,9593	0,9313
25		1	0,9997	0,9992	0,9974	0,9938	0,9869	0,9748	0,9554
26			0,9999	0,9997	0,9987	0,9967	0,9926	0,9848	0,9718
27			1	0,9999	0,9994	0,9983	0,9960	0,9912	0,9827
28				1	0,9997	0,9992	0,9979	0,9950	0,9897
29					0,9999	0,9996	0,9989	0,9973	0,9941
30					1	0,9998	0,9995	0,9986	0,9967
31						0,9999	0,9998	0,9993	0,9982
32						1	0,9999	0,9996	0,9990
33							1	0,9998	0,9995
34								0,9999	0,9998
35								1	0,9999
36									1

TABLE A.5 FONCTION DE RÉPARTITION DE LA LOI NORMALE RÉDUITE
(Probabilité de trouver une valeur inférieure à u)

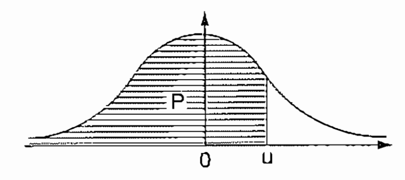


u	0,00	0.01	0,02	0.03	0,04	0,05	0,06	0,07	0.08	0,09
0,0	0.5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0.5279	0,5319	0.5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0.5910	0,5948	0,5987	0.6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0.4	0.6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0.5	0,6915	0.6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0.7486	0,7517	0,7549
0.7	0,7580	0,7611	0,7642	0,7673	0,7704	0.7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0.7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0.8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0.8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0.9222	0,9236	0,9251	0,9265	0,9279	0,9292	0.9306	0,9319
1,5	0,9332	0,9345	0,9357	0.9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0.9554	0,9564	0,9573	0.9582	0,9591	0,9599	0,9608	0.9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0.9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0.9821	0,9826	0,9830	0.9834	0,9838	0,9842	0,9846	0.9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0.9893	0,9896	0,9898	0.9901	0,9904	0,9906	0.9909	0.9911	0.9913	0.9916
2,4	0.9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2.5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0.9952
2,6	0,9953	0.9955	0.9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2.7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0.9974
2,8	0,9974	0.9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de u

u	3,0	3,1	3.2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,99865	0,99904	0.99931	0,99952	0,99966	0,99976	0.999841	0.999928	0.999968	0,999997

Table A.5 *bis* Fractiles de la loi normale réduite



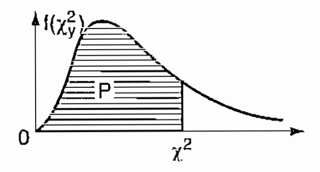
P	0.000	0,001	0,002	0.003	0.004	0,005	0,006	0,007	0.008	0.009	0,010	
0,00	∞	3,0902	2,8782	2,7478	2,6521	2,5758	2,5121	2,4573	2,4089	2,3656	2,3263	0,99
0,01	2,3263	2,2904	2,2571	2,2262	2,1973	2,1701	2,1444	2,1201	2,0969	2,0749	2,0537	0,98
0,02	2,0537	2,0335	2,0141	1,9954	1,9774	1,9600	1,9431	1,9268	1,9110	1,8957	1,8808	0,97
0,03	1,8808	1,8663	1,8522	1,8384	1,8250	1,8119	1,7991	1,7866	1,7744	1,7624	1,7507	0,96
0,04	1,7507	1,7392	1,7279	1,7169	1,7060	1,6954	1,6849	1,6747	1,6646	1,6546	1,6449	0,95
0,05	1,6449	1,6352	1,6258	1,6164	1,6072	1,5982	1,5893	1,5805	1,5718	1,5632	1,5548	0,94
0,06	1,5548	1,5464	1,5382	1,5301	1,5220	1,5141	1,5063	1,4985	1,4909	1,4833	1,4758	0,93
0,07	1,4758	1,4684	1,4611	1,4538	1,4466	1,4395	1,4325	1,4255	1,4187	1,4118	1,4051	0,92
0,08	1,4051	1,3984	1,3917	1,3852	1,3787	1,3722	1,3658	1,3595	1,3532	1,3469	1,3408	0,91
0,09	1,3408	1,3346	1,3285	1,3225	1,3165	1,3106	1,3047	1,2988	1,2930	1,2873	1,2816	0,90
0,10	1,2816	1,2759	1,2702	1,2646	1,2591	1,2536	1,2481	1,2426	1,2372	1,2319	1,2265	0,89
0,11	1,2265	1,2212	1,2160	1,2107	1,2055	1,2004	1,1952	1,1901	1,1850	1,1800	1,1750	0,88
0,12	1,1750	1,1700	1,1650	1,1601	1,1552	1,1503	1,1455	1,1407	1,1359	1,1311	1,1264	0,87
0,13	1,1264	1,1217	1,1170	1,1123	1,1077	1,1031	1,0985	1,0939	1,0893	1,0848	1,0803	0,86
0,14	1,0803	1,0758	1,0714	1,0669	1,0625	1,0581	1,0537	1,0494	1,0450	1,0407	1,0364	0,85
0,15	1,0364	1,0322	1,0279	1,0237	1,0194	1,0152	1,0110	1,0069	1,0027	0,9986	0,9945	0,84
0,16	0,9945	0,9904	0,9863	0,9822	0,9782	0,9741	0,9701	0,9661	0,9621	0,9581	0,9542	0,83
0,17	0,9542	0,9502	0,9463	0,9424	0,9385	0,9346	0,9307	0,9269	0,9230	0,9192	0,9154	0,82
0,18	0,9154	0,9116	0,9078	0,9040	0,9002	0,8965	0,8927	0,8890	0,8853	0,8816	0,8779	0,81
0,19	0,8779	0,8742	0,8705	0,8669	0,8633	0,8596	0,8560	0,8524	0,8488	0,8452	0,8416	0,80
0,20	0,8416	0,8381	0,8345	0,8310	0,8274	0,8239	0,8204	0,8169	0,8134	0,8099	0,8064	0,79
0,21	0,8064	0,8030	0,7995	0,7961	0,7926	0,7892	0,7858	0,7824	0,7790	0,7756	0,7722	0,78
0,22	0,7722	0,7688	0,7655	0,7621	0,7588	0,7554	0,7521	0,7488	0,7454	0,7421	0,7388	0,77
0,23	0,7388	0,7356	0,7323	0,7290	0,7257	0,7225	0,7192	0,7160	0,7128	0,7095	0,7063	0,76
0,24	0,7063	0,7031	0,6999	0,6967	0,6935	0,6903	0,6871	0,6840	0,6808	0,6776	0,6745	0,75
0,25	0,6745	0,6713	0,6682	0,6651	0,6620	0,6588	0,6557	0,6526	0,6495	0,6464	0,6433	0,74
0,26	0,6433	0,6403	0,6372	0,6341	0,6311	0,6280	0,6250	0,6219	0,6189	0,6158	0,6128	0,73
0,27	0,6128	0,6098	0,6068	0,6038	0,6008	0,5978	0,5948	0,5918	0,5888	0,5858	0,5828	0,72
0,28	0,5828	0,5799	0,5769	0,5740	0,5710	0,5681	0,5651	0,5622	0,5592	0,5563	0,5534	0,71
0,29	0,5534	0,5505	0,5476	0,5446	0,5417	0,5388	0,5359	0,5330	0,5302	0,5273	0,5244	0,70
0,30	0,5244	0,5215	0,5187	0,5158	0,5129	0,5101	0,5072	0,5044	0,5015	0,4987	0,4959	0,69
0,31	0,4959	0,4930	0,4902	0,4874	0,4845	0,4817	0,4789	0,4761	0,4733	0,4705	0,4677	0,68
0,32	0,4677	0,4649	0,4621	0,4593	0,4565	0,4538	0,4510	0,4482	0,4454	0,4427	0,4399	0,67
0,33	0,4399	0,4372	0,4344	0,4316	0,4289	0,4261	0,4234	0,4207	0,4179	0,4152	0,4125	0,66
0,34	0,4125	0,4097	0,4070	0,4043	0,4016	0,3989	0,3961	0,3934	0,3907	0,3880	0,3853	0,65
0,35	0,3853	0,3826	0,3799	0,3772	0,3745	0,3719	0,3692	0,3665	0,3638	0,3611	0,3585	0,64
0,36	0,3585	0,3558	0,3531	0,3505	0,3478	0,3451	0,3425	0,3398	0,3372	0,3345	0,3319	0,63
0,37	0,3319	0,3292	0,3266	0,3239	0,3213	0,3186	0,3160	0,3134	0,3107	0,3081	0,3055	0,62
0,38	0,3055	0,3029	0,3002	0,2976	0,2950	0,2924	0,2898	0,2871	0,2845	0,2819	0,2793	0,61
0,39	0,2793	0,2767	0,2741	0,2715	0,2689	0,2663	0,2637	0,2611	0,2585	0,2559	0,2533	0,60
0,40	0,2533	0,2508	0,2482	0,2456	0,2430	0,2404	0,2378	0,2353	0,2327	0,2301	0,2275	0,59
0,41	0,2275	0,2250	0,2224	0,2198	0,2173	0,2147	0,2121	0,2096	0,2070	0,2045	0,2019	0,58
0,42	0,2019	0,1993	0,1968	0,1942	0,1917	0,1891	0,1866	0,1840	0,1815	0,1789	0,1764	0,57
0,43	0,1764	0,1738	0,1713	0,1687	0,1662	0,1637	0,1611	0,1586	0,1560	0,1535	0,1510	0,56
0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257	0,55
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004	0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753	0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502	0,52
0,48	0,0502	0,0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251	0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000	0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

Grandes valeurs de *u*

P	0,9999	0,99999	0,999999	0,9999999	0,99999999	0,999999999
u	3,7190	4,2649	4,7534	5,1993	5,6120	5,9978

N.B. Si $P < 0{,}5$, u est négatif.

TABLE A.6 FRACTILES DE LA LOI DU χ^2. ν NOMBRE DE DEGRÉS DE LIBERTÉ



ν \ P	0,00050	0,0010	0,0050	0,010	0,0250	0,050	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,950	0,9750	0,990	0,9950	0,9990	0,99950
1	0,000000393	0,00000157	0,0000393	0,000157	0,000982	0,00393	0,0158	0,0642	0,148	0,275	0,455	0,708	1,074	1,642	2,706	3,841	5,024	6,635	7,879	10,828	12,116
2	0,00100	0,00200	0,0100	0,0201	0,0506	0,103	0,211	0,446	0,713	1,022	1,386	1,833	2,408	3,219	4,605	5,991	7,378	9,210	10,597	13,816	15,202
3	0,0153	0,0243	0,0717	0,115	0,216	0,352	0,584	1,005	1,424	1,869	2,366	2,946	3,665	4,642	6,251	7,815	9,348	11,345	12,838	16,266	17,730
4	0,0639	0,0908	0,207	0,297	0,484	0,711	1,064	1,649	2,195	2,753	3,357	4,045	4,878	5,989	7,779	9,488	11,143	13,277	14,860	18,467	19,998
5	0,158	0,210	0,412	0,554	0,831	1,145	1,610	2,343	3,000	3,655	4,351	5,132	6,064	7,289	9,236	11,070	12,832	15,086	16,750	20,515	22,105
6	0,299	0,381	0,676	0,872	1,237	1,635	2,204	3,070	3,828	4,570	5,348	6,211	7,231	8,558	10,645	12,592	14,449	16,812	18,548	22,458	24,103
7	0,485	0,598	0,989	1,239	1,690	2,167	2,833	3,822	4,671	5,493	6,346	7,283	8,383	9,803	12,017	14,067	16,013	18,475	20,278	24,322	26,018
8	0,710	0,857	1,344	1,646	2,180	2,733	3,490	4,594	5,527	6,423	7,344	8,351	9,524	11,030	13,362	15,507	17,535	20,090	21,955	26,125	27,868
9	0,972	1,153	1,735	2,088	2,700	3,325	4,168	5,380	6,393	7,357	8,343	9,414	10,656	12,242	14,684	16,919	19,023	21,666	23,589	27,877	29,666
10	1,265	1,479	2,156	2,558	3,247	3,940	4,865	6,179	7,267	8,295	9,342	10,473	11,781	13,442	15,987	18,307	20,483	23,209	25,188	29,588	31,419
11	1,587	1,834	2,603	3,053	3,816	4,575	5,578	6,989	8,148	9,237	10,341	11,530	12,899	14,631	17,275	19,675	21,920	24,725	26,757	31,264	33,136
12	1,934	2,214	3,074	3,571	4,404	5,226	6,304	7,807	9,034	10,182	11,340	12,584	14,011	15,812	18,549	21,026	23,336	26,217	28,300	32,909	34,821
13	2,305	2,617	3,565	4,107	5,009	5,892	7,042	8,634	9,926	11,129	12,340	13,636	15,119	16,985	19,812	22,362	24,736	27,688	29,819	34,528	36,478
14	2,697	3,041	4,075	4,660	5,629	6,571	7,790	9,467	10,821	12,079	13,339	14,685	16,222	18,151	21,064	23,685	26,119	29,141	31,319	36,123	38,109
15	3,108	3,483	4,601	5,229	6,262	7,261	8,547	10,307	11,721	13,030	14,339	15,733	17,322	19,311	22,307	24,996	27,488	30,578	32,801	37,697	39,719
16	3,536	3,942	5,142	5,812	6,908	7,962	9,312	11,152	12,624	13,983	15,338	16,780	18,418	20,465	23,542	26,296	28,845	32,000	34,267	39,252	41,308
17	3,980	4,416	5,697	6,408	7,564	8,672	10,085	12,002	13,531	14,937	16,338	17,824	19,511	21,615	24,769	27,587	30,191	33,409	35,718	40,790	42,879
18	4,439	4,905	6,265	7,015	8,231	9,390	10,865	12,857	14,440	15,893	17,338	18,868	20,601	22,760	25,989	28,869	31,526	34,805	37,156	42,312	44,434
19	4,912	5,407	6,844	7,633	8,907	10,117	11,651	13,716	15,352	16,850	18,338	19,910	21,689	23,900	27,204	30,144	32,852	36,191	38,582	43,820	45,973
20	5,398	5,921	7,434	8,260	9,591	10,851	12,443	14,578	16,266	17,809	19,337	20,951	22,775	25,038	28,412	31,410	34,170	37,566	39,997	45,315	47,498
21	5,896	6,447	8,034	8,897	10,283	11,591	13,240	15,445	17,182	18,768	20,337	21,991	23,858	26,171	29,615	32,671	35,479	38,932	41,401	46,797	49,010
22	6,405	6,983	8,643	9,542	10,982	12,338	14,041	16,314	18,101	19,729	21,337	23,031	24,939	27,301	30,813	33,924	36,781	40,289	42,796	48,268	50,511
23	6,924	7,529	9,260	10,196	11,688	13,091	14,848	17,187	19,021	20,690	22,337	24,069	26,018	28,429	32,007	35,172	38,076	41,638	44,181	49,728	52,000
24	7,453	8,085	9,886	10,856	12,401	13,848	15,659	18,062	19,943	21,652	23,337	25,106	27,096	29,553	33,196	36,415	39,364	42,980	45,558	51,179	53,479

TABLE A.6 (*suite*) FRACTILES DE LA LOI DU χ^2. ν NOMBRE DE DEGRÉS DE LIBERTÉ

ν \ P	0,00050	0,0010	0,0050	0,010	0,0250	0,050	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,950	0,9750	0,990	0,9950	0,9990	0,99950
25	7,991	8,649	10,520	11,524	13,120	14,611	16,473	18,940	20,867	22,616	24,337	26,143	28,172	30,675	34,382	37,652	40,646	44,314	46,928	52,620	54,947
26	8,538	9,222	11,160	12,198	13,844	15,379	17,292	19,820	21,792	23,579	25,336	27,179	29,246	31,795	35,563	38,885	41,923	45,642	48,290	54,052	56,407
27	9,093	9,803	11,808	12,879	14,573	16,151	18,114	20,703	22,719	24,544	26,336	28,214	30,319	32,912	36,741	40,113	43,194	46,963	49,645	55,476	57,858
28	9,656	10,391	12,461	13,565	15,308	16,928	18,939	21,588	23,647	25,509	27,336	29,249	31,391	34,027	37,916	41,337	44,461	48,278	50,993	56,892	59,300
29	10,227	10,986	13,121	14,256	16,047	17,708	19,768	22,475	24,577	26,475	28,336	30,283	32,461	35,139	39,087	42,557	45,722	49,588	52,336	58,302	60,734
30	10,804	11,588	13,787	14,953	16,791	18,493	20,599	23,364	25,508	27,442	29,336	31,316	33,530	36,250	40,256	43,773	46,979	50,892	53,672	59,703	62,161
31	11,389	12,196	14,458	15,655	17,539	19,281	21,434	24,255	26,440	28,409	30,336	32,349	34,598	37,359	41,422	44,985	48,232	52,191	55,003	61,098	63,582
32	11,979	12,811	15,134	16,362	18,291	20,072	22,271	25,148	27,373	29,376	31,336	33,381	35,665	38,466	42,585	46,194	49,480	53,486	56,328	62,487	64,995
33	12,576	13,431	15,815	17,073	19,047	20,867	23,110	26,042	28,307	30,344	32,336	34,413	36,731	39,572	43,745	47,400	50,725	54,776	57,648	63,870	66,402
34	13,179	14,057	16,501	17,789	19,806	21,664	23,952	26,938	29,242	31,313	33,336	35,444	37,795	40,676	44,903	48,602	51,966	56,061	58,964	65,247	67,803
35	13,788	14,688	17,192	18,509	20,569	22,465	24,797	27,836	30,178	32,282	34,336	36,475	38,859	41,778	46,059	49,802	53,203	57,342	60,275	66,619	69,198
36	14,401	15,324	17,887	19,233	21,336	23,269	25,643	28,735	31,115	33,252	35,336	37,505	39,922	42,879	47,212	50,998	54,437	58,619	61,581	67,985	70,588
37	15,020	15,965	18,586	19,960	22,106	24,075	26,492	29,635	32,053	34,222	36,336	38,535	40,984	43,978	48,363	52,192	55,668	59,892	62,883	69,346	71,972
38	15,644	16,611	19,289	20,691	22,878	24,884	27,343	30,537	32,992	35,192	37,335	39,564	42,045	45,076	49,513	53,384	56,895	61,162	64,181	70,703	73,351
39	16,273	17,261	19,996	21,426	23,654	25,695	28,196	31,441	33,932	36,163	38,335	40,593	43,105	46,173	50,660	54,572	58,120	62,428	65,476	72,055	74,725
40	16,906	17,916	20,707	22,164	24,433	26,509	29,051	32,345	34,872	37,134	39,335	41,622	44,165	47,269	51,805	55,758	59,342	63,691	66,766	73,402	76,095
41	17,544	18,575	21,421	22,906	25,215	27,326	29,907	33,251	35,813	38,105	40,335	42,651	45,224	48,363	52,949	56,942	60,561	64,950	68,053	74,745	77,459
42	18,186	19,238	22,138	23,650	25,999	28,144	30,765	34,157	36,755	39,077	41,335	43,679	46,282	49,456	54,090	58,124	61,777	66,206	69,336	76,084	78,820
43	18,832	19,905	22,859	24,398	26,785	28,965	31,625	35,065	37,698	40,050	42,335	44,706	47,339	50,548	55,230	59,304	62,990	67,459	70,616	77,418	80,176
44	19,482	20,576	23,584	25,148	27,575	29,787	32,487	35,974	38,641	41,022	43,335	45,734	48,396	51,639	56,369	60,481	64,201	68,709	71,893	78,749	81,528
45	20,136	21,251	24,311	25,901	28,366	30,612	33,350	36,884	39,585	41,995	44,335	46,761	49,452	52,729	57,505	61,656	65,410	69,957	73,166	80,077	82,876
46	20,794	21,929	25,041	26,657	29,160	31,439	34,215	37,795	40,529	42,968	45,335	47,787	50,507	53,818	58,641	62,830	66,617	71,201	74,437	81,400	84,220
47	21,456	22,610	25,774	27,416	29,956	32,268	35,081	38,708	41,474	43,942	46,335	48,814	51,562	54,906	59,774	64,001	67,821	72,443	75,704	82,720	85,560
48	22,121	23,295	26,511	28,177	30,755	33,098	35,949	39,621	42,420	44,915	47,335	49,840	52,616	55,993	60,907	65,171	69,023	73,683	76,969	84,037	86,897
49	22,789	23,983	27,249	28,941	31,555	33,930	36,818	40,534	43,366	45,889	48,335	50,866	53,670	57,079	62,038	66,339	70,222	74,919	78,231	85,350	88,231

TABLE A.6 (*suite*) FRACTILES DE LA LOI DU χ^2. ν NOMBRE DE DEGRÉS DE LIBERTÉ

ν \ P	0,00050	0,0010	0,0050	0,010	0,0250	0,050	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,950	0,9750	0,990	0,9950	0,9990	0,99950
50	23,461	24,674	27,991	29,707	32,357	34,764	37,689	41,449	44,313	46,864	49,335	51,892	54,723	58,164	63,167	67,505	71,420	76,154	79,490	86,661	89.561
51	24,136	25,368	28,735	30,475	33,162	35,600	38,560	42,365	45,261	47,838	50,335	52,917	55,775	59,248	64,295	68,669	72,616	77,386	80,747	87,968	90,887
52	24,814	26,065	29,481	31,246	33,968	36,437	39,433	43,281	46,209	48,813	51,335	53,942	56,827	60,332	65,422	69,832	73,810	78,616	82,001	89,272	92,211
53	25.495	26,765	30,230	32,018	34,776	37,276	40,308	44,199	47,157	49,788	52,335	54,967	57,879	61,414	66.548	70,993	75,002	79,843	83,253	90,573	93,532
54	26.179	27,468	30,981	32,793	35,586	38,116	41,183	45,117	48,106	50,764	53,335	55,992	58,930	62,496	67,673	72,153	76,192	81,069	84.502	91,872	94,849
55	26,866	28,173	31,735	33,570	36,398	38,958	42,060	46,036	49,056	51,739	54,335	57,016	59,980	63,577	68,796	73,311	77,380	82,292	85,749	93,167	96,163
56	27,556	28,881	32,490	34,350	37,212	39,801	42,937	46,955	50,005	52,715	55,335	58,040	61,031	64,658	69,918	74,468	78.567	83,513	86,994	94,460	97,475
57	28,248	29,592	33,248	35,131	38,027	40,646	43,816	47,876	50,956	53,691	56,335	59,064	62,080	65,737	71,040	75.624	79,752	84,733	88,236	95,751	98,784
58	28,943	30,305	34,008	35,913	38,844	41,492	44,696	48,797	51,906	54,667	57,335	60,088	63,129	66,816	72,160	76,778	80,936	85,950	89,477	97,039	100,090
59	29,640	31,021	34,771	36,698	39,662	42,339	45,577	49,718	52,857	55,643	58,335	61,111	64,178	67,894	73,279	77,931	82,117	87.166	90,715	98,324	101,394
60	30,340	31,739	35,535	37,485	40,482	43,188	46,459	50,641	53,809	56,620	59,335	62.135	65,226	68,972	74,397	79,082	83,298	88,379	91,952	99,607	102,695
61	31,043	32,459	36,301	38,273	41,303	44,038	47,342	51.564	54,761	57,597	60,335	63,158	66,274	70,049	75,514	80,232	84,476	89,591	93,186	100,888	103,993
62	31,748	33,181	37,068	39,063	42,126	44,889	48,226	52,487	55,714	58,574	61,335	64,181	67,322	71,125	76,630	81,381	85,654	90,802	94,419	102,166	105.289
63	32,455	33,906	37,838	39,855	42,950	45,741	49,111	53,411	56,666	59,551	62,335	65,204	68,369	72,201	77,745	82,529	86.830	92,010	95,649	103,442	106,583
64	33,165	34,633	38,610	40,649	43,776	46,595	49,996	54,336	57,619	60,528	63,335	66,226	69,416	73,276	78,860	83,675	88,004	93,217	96,878	104,716	107,874
65	33,877	35,362	39,383	41,444	44,603	47,450	50,883	55,262	58,573	61,506	64,335	67,249	70,462	74,351	79,973	84,821	89,177	94,422	98,105	105,988	109,164
66	34,591	36,093	40,158	42,240	45,431	48,305	51,770	56,188	59,527	62,484	65,335	68,271	71,508	75,425	81,086	85,965	90,349	95,626	99,330	107,258	110,451
67	35,307	36,826	40,935	43,038	46,261	49,162	52,659	57,115	60,481	63,461	66,335	69,293	72,554	76,498	82,197	87,108	91,519	96,828	100,554	108,525	111,735
68	36,025	37,561	41,713	43,838	47,092	50,020	53,548	58,042	61,436	64,440	67,334	70,315	73,600	77,571	83,308	88,250	92,688	98,028	101,776	109,791	113,018
69	36,745	38,298	42,494	44,639	47,924	50,879	54,438	58,970	62,391	65,418	68,334	71,337	74,645	78,643	84,418	89,391	93,856	99,227	102,996	111,055	114,299
70	37,467	39,036	43.275	45,442	48,758	51,739	55,329	59,898	63,346	66,396	69,334	72,358	75,689	79,715	85,527	90,531	95,023	100,425	104,215	112,317	115,577
71	38,192	39.777	44,058	46,246	49.592	52,600	56,221	60,827	64,302	67,375	70,334	73,380	76,734	80,786	86,635	91,670	96,189	101,621	105,432	113,577	116,854
72	38,918	40,520	44,843	47,051	50,428	53,462	57,113	61,756	65,258	68,353	71,334	74,401	77,778	81,857	87,743	92,808	97,353	102,816	106,648	114,835	118,129
73	39,646	41,264	45,629	47,858	51,265	54,325	58,006	62,686	66,214	69,332	72,334	75,422	78,822	82,927	88,850	93,945	98,516	104,010	107,862	116,091	119,402
74	40,376	42,010	46,417	48,666	52,103	55,189	58,900	63,616	67,170	70,311	73,334	76,443	79,865	83,997	89,956	95,081	99,678	105,202	109,074	117.346	120,673

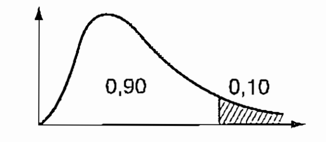
TABLE A.6 *(suite et fin)* FRACTILES DE LA LOI DU χ^2. ν NOMBRE DE DEGRÉS DE LIBERTÉ

ν \ P	0,00050	0,0010	0,0050	0,010	0,0250	0,050	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,950	0,9750	0,990	0,9950	0,9990	0,99950
75	41,107	42,757	47,206	49,475	52,942	56,054	59,795	64,547	68,127	71,290	74,334	77,464	80,908	85,066	91,061	96,217	100,839	106,393	110,286	118,599	121,942
76	41,841	43,506	47,998	50,286	53,782	56,920	60,690	65,478	69,084	72,270	75,334	78,485	81,951	86,135	92,166	97,351	101,999	107,583	111,495	119,851	123,209
77	42,576	44,257	48,788	51,097	54,623	57,786	61,586	66,409	70,042	73,249	76,334	79,505	82,994	87,203	93,270	98,484	103,158	108,771	112,704	121,100	124,475
78	43,313	45,010	49,582	51,910	55,466	58,654	62,483	67,341	70,999	74,228	77,334	80,526	84,036	88,271	94,374	99,617	104,316	109,958	113,911	122,348	125,739
79	44,051	45,764	50,376	52,725	56,309	59,522	63,380	68,274	71,957	75,208	78,334	81,546	85,078	89,338	95,476	100,749	105,473	111,144	115,117	123,594	127,001
80	44,791	46,520	51,172	53,540	57,153	60,391	64,278	69,207	72,915	76,188	79,334	82,566	86,120	90,405	96,578	101,879	106,629	112,329	116,321	124,839	128,261
81	45,533	47,277	51,969	54,357	57,998	61,261	65,176	70,140	73,874	77,168	80,334	83,586	87,161	91,472	97,680	103,009	107,783	113,512	117,524	126,083	129,520
82	46,276	48,036	52,767	55,174	58,845	62,132	66,076	71,074	74,833	78,148	81,334	84,606	88,202	92,538	98,780	104,139	108,937	114,695	118,726	127,324	130,777
83	47,021	48,796	53,567	55,993	59,692	63,004	66,976	72,008	75,792	79,128	82,334	85,626	89,243	93,604	99,880	105,267	110,090	115,876	119,927	128,565	132,033
84	47,767	49,557	54,368	56,813	60,540	63,876	67,876	72,943	76,751	80,108	83,334	86,646	90,284	94,669	100,980	106,395	111,242	117,057	121,126	129,804	133,287
85	48,515	50,320	55,170	57,634	61,389	64,749	68,777	73,878	77,710	81,089	84,334	87,665	91,325	95,734	102,079	107,522	112,393	118,236	122,325	131,041	134,540
86	49,264	51,085	55,973	58,456	62,239	65,623	69,679	74,813	78,670	82,069	85,334	88,685	92,365	96,799	103,177	108,648	113,544	119,414	123,522	132,277	135,792
87	50,015	51,850	56,777	59,279	63,089	66,498	70,581	75,749	79,630	83,050	86,334	89,704	93,405	97,863	104,275	109,773	114,693	120,591	124,718	133,512	137,042
88	50,767	52,617	57,582	60,103	63,941	67,373	71,484	76,685	80,590	84,031	87,334	90,723	94,445	98,927	105,372	110,898	115,841	121,767	125,912	134,745	138,290
89	51,521	53,386	58,389	60,928	64,793	68,249	72,387	77,622	81,550	85,012	88,334	91,742	95,484	99,991	106,469	112,022	116,989	122,942	127,106	135,977	139,537
90	52,276	54,155	59,196	61,754	65,647	69,126	73,291	78,558	82,511	85,993	89,334	92,761	96,524	101,054	107,565	113,145	118,136	124,116	128,299	137,208	140,783
91	53,032	54,926	60,005	62,581	66,501	70,003	74,196	79,496	83,472	86,974	90,334	93,780	97,563	102,116	108,661	114,268	119,282	125,289	129,491	138,438	142,027
92	53,790	55,698	60,815	63,409	67,356	70,882	75,101	80,433	84,433	87,955	91,334	94,799	98,602	103,179	109,756	115,390	120,427	126,462	130,681	139,666	143,270
93	54,549	56,471	61,625	64,238	68,211	71,760	76,006	81,371	85,394	88,936	92,334	95,818	99,641	104,242	110,850	116,511	121,571	127,633	131,871	140,893	144,511
94	55,309	57,246	62,437	65,068	69,068	72,640	76,912	82,309	86,356	89,917	93,334	96,836	100,679	105,303	111,944	117,632	122,715	128,803	133,059	142,119	145,751
95	56,070	58,022	63,250	65,898	69,925	73,520	77,818	83,248	87,317	90,899	94,334	97,855	101,717	106,364	113,038	118,752	123,858	129,973	134,247	143,343	146,990
96	56,833	58,799	64,063	66,730	70,783	74,400	78,725	84,187	88,279	91,881	95,334	98,873	102,755	107,425	114,131	119,871	125,000	131,141	135,433	144,567	148,228
97	57,597	59,577	64,878	67,562	71,642	75,282	79,633	85,126	89,241	92,862	96,334	99,892	103,793	108,486	115,223	120,990	126,141	132,309	136,619	145,789	149,464
98	58,362	60,356	65,694	68,396	72,501	76,164	80,541	86,065	90,204	93,844	97,334	100,910	104,831	109,547	116,315	122,108	127,282	133,476	137,803	147,010	150,699
99	59,128	61,136	66,510	69,230	73,361	77,046	81,449	87,005	91,166	94,826	98,334	101,928	105,868	110,607	117,406	123,225	128,422	134,642	138,987	148,230	151,934
100	59,897	61,919	67,328	70,065	74,222	77,930	82,358	87,945	92,129	95,808	99,334	102,946	106,906	111,667	118,498	124,342	129,561	135,806	140,169	149,448	153,165

Pour $\nu > 100$ on utilisera l'une des deux approximations suivantes, la seconde étant de loin la meilleure :

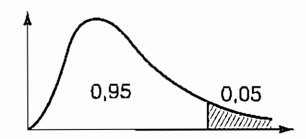
$$a)\ \sqrt{2\chi^2_\nu} - \sqrt{2\nu - 1} \simeq U \qquad b)\ \left[\left(\frac{\chi^2_\nu}{\nu}\right)^{1/3} + \frac{2}{9\nu} - 1\right]\sqrt{\frac{9\nu}{2}} \simeq U$$

TABLE A.7 VALEURS f DE LA VARIABLE DE FISHER-SNEDECOR $F(\nu_1 ; \nu_2)$ AYANT LA PROBABILITÉ 0.10 D'ÊTRE DÉPASSÉES



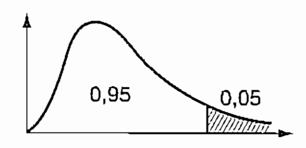
ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.50	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.80	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

TABLE A.7 (*suite*) VALEURS f DE LA VARIABLE DE FISHER-SNEDECOR $F(\nu_1 ; \nu_2)$ AYANT LA PROBABILITÉ 0.05 D'ÊTRE DÉPASSÉES



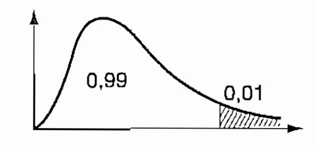
ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	161	200	216	225	230	234	237	239	241	242	243	244	245	245	246	246	247	247
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96
10	4.90	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13	2.11	2.10
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.23	2.20	2.18	2.15	2.13	2.11	2.09	2.07
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.21	2.18	2.15	2.13	2.11	2.09	2.07	2.05
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.07	2.05	2.04
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	2.05	2.03	2.02
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	2.04	2.02	2.00
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	2.02	2.00	1.99
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	2.01	1.99	1.97
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.04	2.01	1.99	1.97	1.95	1.94
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05	2.02	1.99	1.97	1.95	1.93	1.92
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03	2.00	1.98	1.95	1.93	1.92	1.90
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.99	1.96	1.94	1.92	1.90	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90	1.89	1.87
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.03	1.99	1.96	1.93	1.91	1.89	1.87	1.86
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.95	1.92	1.90	1.88	1.86	1.84
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04	2.00	1.97	1.94	1.91	1.89	1.87	1.85	1.83
48	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03	1.99	1.96	1.93	1.90	1.88	1.86	1.84	1.82
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.85	1.83	1.81
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01	1.97	1.93	1.90	1.88	1.85	1.83	1.81	1.79
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.82	1.80	1.78
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.03	1.98	1.94	1.90	1.87	1.85	1.82	1.80	1.78	1.76
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84	1.81	1.79	1.77	1.75
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.84	1.82	1.79	1.77	1.75	1.73
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86	1.83	1.80	1.78	1.76	1.74	1.72
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.75	1.73	1.71
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.96	1.91	1.87	1.83	1.80	1.77	1.75	1.72	1.70	1.69
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.79	1.76	1.73	1.71	1.69	1.67
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80	1.77	1.74	1.72	1.69	1.67	1.66
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.82	1.78	1.75	1.72	1.70	1.68	1.66	1.64
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.77	1.74	1.71	1.69	1.66	1.64	1.62
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.80	1.76	1.73	1.70	1.68	1.65	1.63	1.61
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.72	1.69	1.67	1.64	1.62	1.60

TABLE A.7 (*suite*) VALEURS f DE LA VARIABLE DE FISHER-SNEDECOR $F(\nu_1\ ;\ \nu_2)$ AYANT LA PROBABILITÉ 0.05 D'ÊTRE DÉPASSÉES



ν_2 \ ν_1	19	20	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞
1	248	248	249	249	249	250	250	251	251	251	252	252	252	253	254	254	254
2	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	8.67	8.66	8.65	8.64	8.63	8.62	8.62	8.60	8.59	8.59	8.58	8.57	8.56	8.55	8.54	8.53	8.53
4	5.81	5.80	5.79	5.77	5.76	5.75	5.75	5.73	5.72	5.71	5.70	5.69	5.67	5.66	5.65	5.64	5.63
5	4.57	4.56	4.54	4.53	4.52	4.50	4.50	4.48	4.46	4.45	4.44	4.43	4.41	4.41	4.39	4.37	4.37
6	3.88	3.87	3.86	3.84	3.83	3.82	3.81	3.79	3.77	3.76	3.75	3.74	3.72	3.71	3.69	3.68	3.67
7	3.46	3.44	3.43	3.41	3.40	3.39	3.38	3.36	3.34	3.33	3.32	3.30	3.29	3.27	3.25	3.24	3.23
8	3.16	3.15	3.13	3.12	3.10	3.09	3.08	3.06	3.04	3.03	3.02	3.01	2.99	2.97	2.95	2.94	2.93
9	2.95	2.94	2.92	2.90	2.89	2.87	2.86	2.84	2.83	2.81	2.80	2.79	2.77	2.76	2.73	2.72	2.71
10	2.78	2.77	2.75	2.74	2.72	2.71	2.70	2.68	2.66	2.65	2.64	2.62	2.60	2.59	2.56	2.55	2.54
11	2.66	2.65	2.63	2.61	2.59	2.58	2.57	2.55	2.53	2.52	2.51	2.49	2.47	2.46	2.43	2.42	2.40
12	2.56	2.54	2.52	2.51	2.49	2.48	2.47	2.44	2.43	2.41	2.40	2.38	2.36	2.35	2.32	2.31	2.30
13	2.47	2.46	2.44	2.42	2.41	2.39	2.38	2.36	2.34	2.33	2.31	2.30	2.27	2.26	2.23	2.22	2.21
14	2.40	2.39	2.37	2.35	2.33	2.32	2.31	2.28	2.27	2.25	2.24	2.22	2.20	2.19	2.16	2.14	2.13
15	2.34	2.33	2.31	2.29	2.27	2.26	2.25	2.22	2.20	2.19	2.18	2.16	2.14	2.12	2.10	2.08	2.07
16	2.29	2.28	2.25	2.24	2.22	2.21	2.19	2.17	2.15	2.14	2.12	2.11	2.08	2.07	2.04	2.02	2.01
17	2.24	2.23	2.21	2.19	2.17	2.16	2.15	2.12	2.10	2.09	2.08	2.06	2.03	2.02	1.99	1.97	1.96
18	2.20	2.19	2.17	2.15	2.13	2.12	2.11	2.08	2.06	2.05	2.04	2.02	1.99	1.98	1.95	1.93	1.92
19	2.17	2.16	2.13	2.11	2.10	2.08	2.07	2.05	2.03	2.01	2.00	1.98	1.96	1.94	1.91	1.89	1.88
20	2.14	2.12	2.10	2.08	2.07	2.05	2.04	2.01	1.99	1.98	1.97	1.95	1.92	1.91	1.88	1.86	1.84
21	2.11	2.10	2.07	2.05	2.04	2.02	2.01	1.98	1.96	1.95	1.94	1.92	1.89	1.88	1.84	1.82	1.81
22	2.08	2.07	2.05	2.03	2.01	2.00	1.98	1.96	1.94	1.92	1.91	1.89	1.86	1.85	1.82	1.80	1.78
23	2.06	2.05	2.02	2.00	1.99	1.97	1.96	1.93	1.91	1.90	1.88	1.86	1.84	1.82	1.79	1.77	1.76
24	2.04	2.03	2.00	1.98	1.97	1.95	1.94	1.91	1.89	1.88	1.86	1.84	1.82	1.80	1.77	1.75	1.73
25	2.02	2.01	1.98	1.96	1.95	1.93	1.92	1.89	1.87	1.86	1.84	1.82	1.80	1.78	1.75	1.73	1.71
26	2.00	1.99	1.97	1.95	1.93	1.91	1.90	1.87	1.85	1.84	1.82	1.80	1.78	1.76	1.73	1.71	1.69
27	1.99	1.97	1.95	1.93	1.91	1.90	1.88	1.86	1.84	1.82	1.81	1.79	1.76	1.74	1.71	1.69	1.67
28	1.97	1.96	1.93	1.91	1.90	1.88	1.87	1.84	1.82	1.80	1.79	1.77	1.74	1.73	1.69	1.67	1.65
29	1.96	1.94	1.92	1.90	1.88	1.87	1.85	1.83	1.81	1.79	1.77	1.75	1.73	1.71	1.67	1.65	1.64
30	1.95	1.93	1.91	1.89	1.87	1.85	1.84	1.81	1.79	1.77	1.76	1.74	1.71	1.70	1.66	1.64	1.62
32	1.92	1.91	1.88	1.86	1.85	1.83	1.82	1.79	1.77	1.75	1.74	1.71	1.69	1.67	1.63	1.61	1.59
34	1.90	1.89	1.86	1.84	1.82	1.80	1.80	1.77	1.75	1.73	1.71	1.69	1.66	1.65	1.61	1.59	1.57
36	1.88	1.87	1.85	1.82	1.81	1.79	1.78	1.75	1.73	1.71	1.69	1.67	1.64	1.62	1.59	1.56	1.55
38	1.87	1.85	1.83	1.81	1.79	1.77	1.76	1.73	1.71	1.69	1.68	1.65	1.62	1.61	1.57	1.54	1.53
40	1.85	1.84	1.81	1.79	1.77	1.76	1.74	1.72	1.69	1.67	1.66	1.64	1.61	1.59	1.55	1.53	1.51
42	1.84	1.83	1.80	1.78	1.76	1.74	1.73	1.70	1.68	1.66	1.65	1.62	1.59	1.57	1.53	1.51	1.49
44	1.83	1.81	1.79	1.77	1.75	1.73	1.72	1.69	1.67	1.65	1.63	1.61	1.58	1.56	1.52	1.49	1.48
46	1.82	1.80	1.78	1.76	1.74	1.72	1.71	1.68	1.65	1.64	1.62	1.60	1.57	1.55	1.51	1.48	1.46
48	1.81	1.79	1.77	1.75	1.73	1.71	1.70	1.67	1.64	1.62	1.61	1.59	1.56	1.54	1.49	1.47	1.45
50	1.80	1.78	1.76	1.74	1.72	1.70	1.69	1.66	1.63	1.61	1.60	1.58	1.54	1.52	1.48	1.46	1.44
55	1.78	1.76	1.74	1.72	1.70	1.68	1.67	1.64	1.61	1.59	1.58	1.55	1.52	1.50	1.46	1.43	1.41
60	1.76	1.75	1.72	1.70	1.68	1.66	1.65	1.62	1.59	1.57	1.56	1.53	1.50	1.48	1.44	1.41	1.39
65	1.75	1.73	1.71	1.69	1.67	1.65	1.63	1.60	1.58	1.56	1.54	1.52	1.49	1.46	1.42	1.39	1.37
70	1.74	1.72	1.70	1.67	1.65	1.64	1.62	1.59	1.57	1.55	1.53	1.50	1.47	1.45	1.40	1.37	1.35
80	1.72	1.70	1.68	1.65	1.63	1.62	1.60	1.57	1.54	1.52	1.51	1.48	1.45	1.43	1.38	1.35	1.32
90	1.70	1.69	1.66	1.64	1.62	1.60	1.59	1.55	1.53	1.51	1.49	1.46	1.43	1.41	1.36	1.32	1.30
100	1.69	1.68	1.65	1.63	1.61	1.59	1.57	1.54	1.52	1.49	1.48	1.45	1.41	1.39	1.34	1.31	1.28
125	1.67	1.65	1.63	1.60	1.58	1.57	1.55	1.52	1.49	1.47	1.45	1.42	1.39	1.36	1.31	1.27	1.25
150	1.66	1.64	1.61	1.59	1.57	1.55	1.53	1.50	1.48	1.45	1.44	1.41	1.37	1.34	1.29	1.25	1.22
200	1.64	1.62	1.60	1.57	1.55	1.53	1.52	1.48	1.46	1.43	1.41	1.39	1.35	1.32	1.26	1.22	1.19
300	1.62	1.61	1.58	1.55	1.53	1.51	1.50	1.46	1.43	1.41	1.39	1.36	1.32	1.30	1.23	1.19	1.15
500	1.61	1.59	1.56	1.54	1.52	1.50	1.48	1.45	1.42	1.40	1.38	1.34	1.30	1.28	1.21	1.16	1.11
1000	1.60	1.58	1.55	1.53	1.51	1.49	1.47	1.44	1.41	1.38	1.36	1.33	1.29	1.26	1.19	1.13	1.08
∞	1.59	1.57	1.54	1.52	1.50	1.48	1.46	1.42	1.39	1.37	1.35	1.32	1.27	1.24	1.17	1.11	1.00

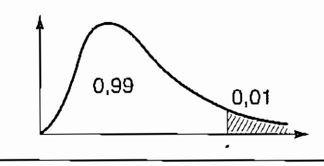
TABLE A.7 (*suite*) VALEURS f DE LA VARIABLE DE FISHER-SNEDECOR $F(\nu_1\ ;\ \nu_2)$ AYANT LA PROBABILITÉ 0.01 D'ÊTRE DÉPASSÉES



(Les valeurs de la première ligne doivent être multipliées par 10)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	*405*	*500*	*540*	*563*	*576*	*586*	*593*	*598*	*602*	*606*	*608*	*611*	*613*	*614*	*616*	*617*	*618*	*619*
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1	27.0	26.9	26.9	26.8	26.8	26.8
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4	14.3	14.2	14.2	14.2	14.1	14.1
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89	9.82	9.77	9.72	9.68	9.64	9.61
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.52	7.48	7.45
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.27	6.24	6.21
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48	5.44	5.41
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.00	4.96	4.92	4.89	4.86
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52	4.49	4.46
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21	4.18	4.15
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97	3.94	3.91
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78	3.75	3.72
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62	3.59	3.56
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49	3.45	3.42
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.37	3.34	3.31
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27	3.24	3.21
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19	3.16	3.13
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.12	3.08	3.05
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05	3.02	2.99
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.99	2.96	2.93
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.94	2.91	2.88
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.89	2.86	2.83
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.85	2.82	2.79
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85	2.81	2.78	2.75
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.82	2.78	2.74	2.72
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78	2.75	2.71	2.68
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75	2.72	2.68	2.65
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73	2.69	2.66	2.63
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.66	2.63	2.60
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.86	2.80	2.74	2.70	2.66	2.62	2.58	2.55
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.82	2.76	2.70	2.66	2.62	2.58	2.55	2.51
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86	2.79	2.72	2.67	2.62	2.58	2.54	2.51	2.48
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.75	2.69	2.64	2.59	2.55	2.51	2.48	2.45
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56	2.52	2.48	2.45	2.42
42	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78	2.70	2.64	2.59	2.54	2.50	2.46	2.43	2.40
44	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75	2.68	2.62	2.56	2.52	2.47	2.44	2.40	2.37
46	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73	2.66	2.60	2.54	2.50	2.45	2.42	2.38	2.35
48	7.19	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.72	2.64	2.58	2.53	2.48	2.44	2.40	2.37	2.33
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70	2.63	2.56	2.51	2.46	2.42	2.38	2.35	2.32
55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53	2.47	2.42	2.38	2.34	2.31	2.28
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39	2.35	2.31	2.28	2.25
65	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61	2.53	2.47	2.42	2.37	2.33	2.29	2.26	2.23
70	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.51	2.45	2.40	2.35	2.31	2.27	2.23	2.20
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.48	2.42	2.36	2.31	2.27	2.23	2.20	2.17
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52	2.45	2.39	2.33	2.29	2.24	2.21	2.17	2.14
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.43	2.37	2.31	2.26	2.22	2.19	2.15	2.12
125	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.66	2.55	2.47	2.39	2.33	2.28	2.23	2.19	2.15	2.11	2.08
150	6.81	4.75	3.92	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.37	2.31	2.25	2.20	2.16	2.12	2.09	2.06
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27	2.22	2.17	2.13	2.09	2.06	2.02
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.31	2.24	2.19	2.14	2.10	2.06	2.03	1.99
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.28	2.22	2.17	2.12	2.07	2.04	2.00	1.97
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.27	2.20	2.15	2.10	2.06	2.02	1.98	1.95
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18	2.13	2.08	2.04	2.00	1.97	1.93

TABLE A.7 (*suite et fin*) VALEURS f DE LA VARIABLE DE FISHER-SNEDECOR $F(\nu_1\ ;\ \nu_2)$ AYANT LA PROBABILITÉ 0.01 D'ÊTRE DÉPASSÉES

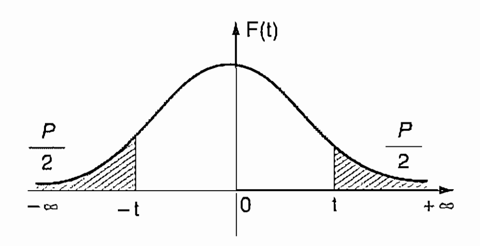


(Les valeurs de la première ligne doivent être multipliées par 10)

ν_2 \ ν_1	19	20	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞
1	*620*	*621*	*622*	*623*	*624*	*625*	*626*	*628*	*629*	*630*	*630*	*631*	*633*	*633*	*635*	*636*	*637*
2	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	26.7	26.7	26.6	26.6	26.6	26.5	26.5	26.5	26.4	26.4	26.4	26.3	26.3	26.2	26.2	26.1	26.1
4	14.0	14.0	14.0	13.9	13.9	13.9	13.8	13.8	13.7	13.7	13.7	13.7	13.6	13.6	13.5	13.5	13.5
5	9.58	9.55	9.51	9.47	9.43	9.40	9.38	9.33	9.29	9.26	9.24	9.20	9.16	9.13	9.08	9.04	9.02
6	7.42	7.40	7.35	7.31	7.28	7.25	7.23	7.18	7.14	7.11	7.09	7.06	7.01	6.99	6.93	6.90	6.88
7	6.18	6.16	6.11	6.07	6.04	6.02	5.99	5.94	5.91	5.88	5.86	5.82	5.78	5.75	5.70	5.67	5.65
8	5.38	5.36	5.32	5.28	5.25	5.22	5.20	5.15	5.12	5.09	5.07	5.03	4.99	4.96	4.91	4.88	4.86
9	4.83	4.81	4.77	4.73	4.70	4.67	4.65	4.60	4.57	4.54	4.52	4.48	4.44	4.42	4.36	4.33	4.31
10	4.43	4.41	4.36	4.33	4.30	4.27	4.25	4.20	4.17	4.14	4.12	4.08	4.04	4.01	3.96	3.93	3.91
11	4.12	4.10	4.06	4.02	3.99	3.96	3.94	3.89	3.86	3.83	3.81	3.78	3.73	3.71	3.66	3.62	3.60
12	3.88	3.86	3.82	3.78	3.75	3.72	3.70	3.65	3.62	3.59	3.57	3.54	3.49	3.47	3.41	3.38	3.36
13	3.69	3.66	3.62	3.59	3.56	3.53	3.51	3.46	3.43	3.40	3.38	3.34	3.30	3.27	3.22	3.19	3.17
14	3.53	3.51	3.46	3.43	3.40	3.37	3.35	3.30	3.27	3.24	3.22	3.18	3.14	3.11	3.06	3.03	3.00
15	3.40	3.37	3.33	3.29	3.26	3.24	3.21	3.17	3.13	3.10	3.08	3.05	3.00	2.98	2.92	2.89	2.87
16	3.28	3.26	3.22	3.18	3.15	3.12	3.10	3.05	3.02	2.99	2.97	2.93	2.89	2.86	2.81	2.78	2.75
17	3.18	3.16	3.12	3.08	3.05	3.03	3.00	2.96	2.92	2.89	2.87	2.83	2.79	2.76	2.71	2.68	2.65
18	3.10	3.08	3.03	3.00	2.97	2.94	2.92	2.87	2.84	2.81	2.78	2.75	2.70	2.68	2.62	2.59	2.57
19	3.03	3.00	2.96	2.92	2.89	2.87	2.84	2.80	2.76	2.73	2.71	2.67	2.63	2.60	2.55	2.51	2.49
20	2.96	2.94	2.90	2.86	2.83	2.80	2.78	2.73	2.69	2.67	2.64	2.61	2.56	2.54	2.48	2.44	2.42
21	2.90	2.88	2.84	2.80	2.77	2.74	2.72	2.67	2.64	2.61	2.58	2.55	2.50	2.48	2.42	2.38	2.36
22	2.85	2.83	2.78	2.75	2.72	2.69	2.67	2.62	2.58	2.55	2.53	2.50	2.45	2.42	2.36	2.33	2.31
23	2.80	2.78	2.74	2.70	2.67	2.64	2.62	2.57	2.54	2.51	2.48	2.45	2.40	2.37	2.32	2.28	2.26
24	2.76	2.74	2.70	2.66	2.63	2.60	2.58	2.53	2.49	2.46	2.44	2.40	2.36	2.33	2.27	2.24	2.21
25	2.72	2.70	2.66	2.62	2.59	2.56	2.54	2.49	2.45	2.42	2.40	2.36	2.32	2.29	2.23	2.19	2.17
26	2.69	2.66	2.62	2.58	2.55	2.53	2.50	2.45	2.42	2.39	2.36	2.33	2.28	2.25	2.19	2.16	2.13
27	2.66	2.63	2.59	2.55	2.52	2.49	2.47	2.42	2.38	2.35	2.33	2.29	2.25	2.22	2.16	2.12	2.10
28	2.63	2.60	2.56	2.52	2.49	2.46	2.44	2.39	2.35	2.32	2.30	2.26	2.22	2.19	2.13	2.09	2.06
29	2.60	2.57	2.53	2.49	2.46	2.44	2.41	2.36	2.33	2.30	2.27	2.23	2.19	2.16	2.10	2.06	2.03
30	2.57	2.55	2.51	2.47	2.44	2.41	2.39	2.34	2.30	2.27	2.25	2.21	2.16	2.13	2.07	2.03	2.01
32	2.53	2.50	2.46	2.42	2.39	2.36	2.34	2.29	2.25	2.22	2.20	2.16	2.11	2.08	2.02	1.98	1.96
34	2.49	2.46	2.42	2.38	2.35	2.32	2.30	2.25	2.21	2.18	2.16	2.12	2.07	2.04	1.98	1.94	1.91
36	2.45	2.43	2.38	2.35	2.32	2.29	2.26	2.21	2.17	2.14	2.12	2.08	2.03	2.00	1.94	1.90	1.87
38	2.42	2.40	2.35	2.32	2.28	2.26	2.23	2.18	2.14	2.11	2.09	2.05	2.00	1.97	1.90	1.86	1.84
40	2.39	2.37	2.33	2.29	2.26	2.23	2.20	2.15	2.11	2.08	2.06	2.02	1.97	1.94	1.87	1.83	1.80
42	2.37	2.34	2.30	2.26	2.23	2.20	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.91	1.85	1.80	1.78
44	2.35	2.32	2.28	2.24	2.21	2.18	2.15	2.10	2.06	2.03	2.01	1.97	1.92	1.89	1.82	1.78	1.75
46	2.33	2.30	2.26	2.22	2.19	2.16	2.13	2.08	2.04	2.01	1.99	1.95	1.90	1.86	1.80	1.75	1.73
48	2.31	2.28	2.24	2.20	2.17	2.14	2.12	2.06	2.02	1.99	1.97	1.93	1.88	1.84	1.78	1.73	1.70
50	2.29	2.27	2.22	2.18	2.15	2.12	2.10	2.05	2.01	1.97	1.95	1.91	1.86	1.82	1.76	1.71	1.68
55	2.25	2.23	2.18	2.15	2.11	2.08	2.06	2.01	1.97	1.93	1.91	1.87	1.81	1.78	1.71	1.67	1.64
60	2.22	2.20	2.15	2.12	2.08	2.05	2.03	1.98	1.94	1.90	1.88	1.84	1.78	1.75	1.68	1.63	1.60
65	2.20	2.17	2.13	2.09	2.06	2.03	2.00	1.95	1.91	1.88	1.85	1.81	1.75	1.72	1.65	1.60	1.57
70	2.18	2.15	2.11	2.07	2.03	2.01	1.98	1.93	1.89	1.85	1.83	1.78	1.73	1.70	1.62	1.57	1.54
80	2.14	2.12	2.07	2.03	2.00	1.97	1.94	1.89	1.85	1.81	1.79	1.75	1.69	1.66	1.58	1.53	1.49
90	2.11	2.09	2.04	2.00	1.97	1.94	1.92	1.86	1.82	1.79	1.76	1.72	1.66	1.62	1.54	1.49	1.46
100	2.09	2.07	2.02	1.98	1.94	1.92	1.89	1.84	1.80	1.76	1.73	1.69	1.63	1.60	1.52	1.47	1.43
125	2.05	2.03	1.98	1.94	1.91	1.88	1.85	1.80	1.76	1.72	1.69	1.65	1.59	1.55	1.47	1.41	1.37
150	2.03	2.00	1.96	1.92	1.88	1.85	1.83	1.77	1.73	1.69	1.66	1.62	1.56	1.52	1.43	1.38	1.33
200	2.00	1.97	1.93	1.89	1.85	1.82	1.79	1.74	1.69	1.66	1.63	1.58	1.52	1.48	1.39	1.33	1.28
300	1.97	1.94	1.89	1.85	1.82	1.79	1.76	1.71	1.66	1.62	1.59	1.55	1.48	1.44	1.35	1.28	1.22
500	1.94	1.92	1.87	1.83	1.79	1.76	1.74	1.68	1.63	1.60	1.56	1.52	1.45	1.41	1.31	1.23	1.16
1000	1.92	1.90	1.85	1.81	1.77	1.74	1.72	1.66	1.61	1.57	1.54	1.50	1.43	1.38	1.28	1.19	1.11
∞	1.90	1.88	1.83	1.79	1.76	1.72	1.70	1.64	1.59	1.55	1.52	1.47	1.40	1.36	1.25	1.15	1.00

TABLE A.8 TABLE DE DISTRIBUTION DE *T* (LOI DE STUDENT)

Valeurs de *T* ayant la probabilité *P* d'être dépassées en valeur absolue



ν \ P	0.90	0,80	0,70	0.60	0.50	0,40	0.30	0.20	0.10	0.05	0,02	0.01	0,001
1	0,158	0.325	0.510	0.727	1,000	1,376	1,963	3,078	6.314	12,706	31.821	63.657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1.386	1,886	2,920	4.303	6,965	9,925	31,598
3	0,137	0.277	0.424	0,584	0,765	0,978	1,250	1.638	2,353	3,182	4.541	5,841	12.929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4.604	8.610
5	0,132	0.267	0.408	0.559	0,727	0,920	1,156	1.476	2.015	2,571	3,365	4,032	6.869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0.549	0,711	0.896	1,119	1.415	1,895	2.365	2,998	3,499	5.408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0.129	0.261	0.398	0,543	0.703	0.883	1,100	1,383	1.833	2.262	2,821	3,250	4,781
10	0,129	0,260	0,397	0.542	0,700	0,879	1.093	1,372	1,812	2,228	2,764	3.169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0.395	0,539	0,695	0.873	1,083	1.356	1.782	2,179	2.681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1.079	1,350	1,771	2,160	2,650	3,012	4.221
14	0,128	0,258	0,393	0.537	0,692	0,868	1,076	1.345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1.074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0.258	0,392	0.535	0,690	0.865	1,071	1,337	1,746	2,120	2,583	2.921	4.015
17	0,128	0,257	0.392	0,534	0.689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0.127	0,257	0,392	0.534	0,688	0,862	1,067	1.330	1.734	2,101	2,552	2.878	3.922
19	0,127	0,257	0,391	0,533	0.688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0.127	0,257	0,391	0.533	0,687	0.860	1,064	1,325	1.725	2,086	2.528	2,845	3,850
21	0.127	0,257	0.391	0.532	0.686	0.859	1,063	1,323	1.721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0.256	0.390	0,532	0,685	0.858	1,060	1,319	1.714	2.069	2.500	2.807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1.059	1,318	1,711	2,064	2,492	2,797	3,745
25	0.127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2.060	2.485	2.787	3.725
26	0,127	0,256	0,390	0.531	0.684	0.856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0.855	1,057	1,314	1,703	2,052	2,473	2.771	3.690
28	0,127	0,256	0,389	0,530	0.683	0,855	1.056	1,313	1,701	2,048	2,467	2.763	3,674
29	0,127	0.256	0,389	0.530	0,683	0,854	1,055	1.311	1,699	2.045	2.462	2.756	3.659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1.684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0.527	0.679	0,848	1,046	1.296	1.671	2,000	2.390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1.041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0.842	1,036	1,282	1,645	1,960	2.326	2,576	3.291

TABLE A.9 VALEURS CRITIQUES DU COEFFICIENT DE CORRÉLATION R D'UN ÉCHANTILLON ISSU D'UNE POPULATION NORMALE OÙ $\rho = 0$

Cette table donne les valeurs r telles que $P(|R| > r) = \alpha$, $\nu = n - 2$ corrélation simple

$\nu = n - 2 - d$ corrélation partielle avec d variables fixées

ν \ α	0,1	0,05	0,01	0,001	ν \ α	0,1	0,05	0,01	0,001
1	0,9877	0,9969	0,9999	1,0000	**25**	0,3233	0,3809	0,4869	0,5974
2	9000	9500	9900	0,9990	26	3172	3739	4785	5880
3	8054	8783	9587	9911	27	3115	3673	4705	5790
4	7293	8114	9172	9741	28	3061	3610	4629	5703
5	0,6694	0,7545	0,8745	0,9509	29	3009	3550	4556	5620
6	6215	7067	8343	9249	**30**	0,2960	0,3494	0,4487	0,5541
7	5822	6664	7977	8983	31	2913	3440	4421	5465
8	5494	6319	7646	8721	32	2869	3388	4357	5392
9	5214	6021	7348	8471	33	2826	3338	4297	5322
10	0,4973	0,5760	0,7079	0,8233	34	2785	3291	4238	5255
11	4762	5529	6835	8010	**35**	0,2746	0,3246	0,4182	0,5189
12	4575	5324	6614	7800	36	2709	3202	4128	5126
13	4409	5139	6411	7604	37	2673	3160	4076	5066
14	4259	4973	6226	7419	38	2638	3120	4026	5007
15	0,4124	0,4821	0,6055	0,7247	39	2605	3081	3978	4951
16	4000	4683	5897	7084	**40**	0,2573	0,3044	0,3932	0,4896
17	3887	4555	5751	6932	41	2542	3008	3887	4843
18	3783	4438	5614	6788	42	2512	2973	3843	4792
19	3687	4329	5487	6652	43	2483	2940	3802	4742
20	0,3598	0,4227	0,5368	0,6524	44	2455	2907	3761	4694
21	3515	4132	5256	6402	**45**	0,2428	0,2875	0,3721	0,4647
22	3438	4044	5151	6287	46	2403	2845	3683	4602
23	3365	3961	5052	6177	47	2377	2816	3646	4558
24	3297	3882	4958	6073	48	2353	2787	3610	4515

TABLE A.9 *(suite)* VALEURS CRITIQUES DU COEFFICIENT DE CORRÉLATION R D'UN ÉCHANTILLON ISSU D'UNE POPULATION NORMALE OÙ $\rho = 0$
Cette table donne les valeurs r telles que $P(|R| > r) = \alpha$, $\nu = n - 2$ corrélation simple
$\nu = n - 2 - d$ corrélation partielle avec d variables fixées

ν \ α	0,1	0,05	0,01	0,001	ν \ α	0,1	0,05	0,01	0,001
49	2329	2759	3575	4473	**75**	0,1889	0,2242	0,2919	0,3678
50	0,2306	0,2732	0,3541	0,4433	76	1876	2227	2900	3655
51	2284	2706	3509	4393	77	1864	2213	2882	3633
52	2262	2681	3477	4355	78	1852	2199	2864	3611
53	2241	2656	3445	4317	79	1841	2185	2847	3590
54	2221	2632	3415	4281	**80**	0,1829	0,2172	0,2830	0,3569
55	0,2201	0,2609	0,3385	0,4245	81	1818	2159	2813	3548
56	2181	2586	3357	4210	82	1807	2146	2796	3527
57	2162	2564	3329	4176	83	1796	2133	2780	3507
58	2144	2542	3301	4143	84	1786	2120	2764	3488
59	2126	2521	3274	4111	**85**	0,1775	0,2108	0,2748	0,3468
60	0,2108	0,2500	0,3248	0,4079	86	1765	2096	2733	3449
61	2091	2480	3223	4048	87	1755	2084	2717	3430
62	2075	2461	3198	4018	88	1745	2072	2702	3412
63	2058	2442	3174	3988	89	1735	2061	2688	3394
64	2042	2423	3150	3959	**90**	0,1726	0,2050	0,2673	0,3376
65	0,2027	0,2405	0,3127	0,3931	91	1716	2039	2659	3358
66	2012	2387	3104	3904	92	1707	2028	2645	3341
67	1997	2369	3081	3877	93	1698	2017	2631	3324
68	1982	2352	3060	3850	94	1689	2006	2617	3307
69	1968	2335	3038	3824	**95**	0,1680	0,1996	0,2604	0,3291
70	0,1954	0,2319	0,3017	0,3798	96	1671	1986	2591	3274
71	1940	2303	2997	3773	97	1663	1976	2578	3258
72	1927	2287	2977	3749	98	1654	1966	2565	3242
73	1914	2272	2957	3725	99	1646	1956	2552	3227
74	1901	2257	2938	3701	**100**	0,1638	0,1946	0,2540	0,3211

TABLE A.9 *(suite)* VALEURS CRITIQUES DU COEFFICIENT DE CORRÉLATION R D'UN ÉCHANTILLON ISSU D'UNE POPULATION NORMALE OÙ $\rho = 0$

Cette table donne les valeurs r telles que $P(|R| > r) = \alpha$, $\nu = n - 2$ corrélation simple

$\nu = n - 2 - d$ corrélation partielle avec d variables fixées

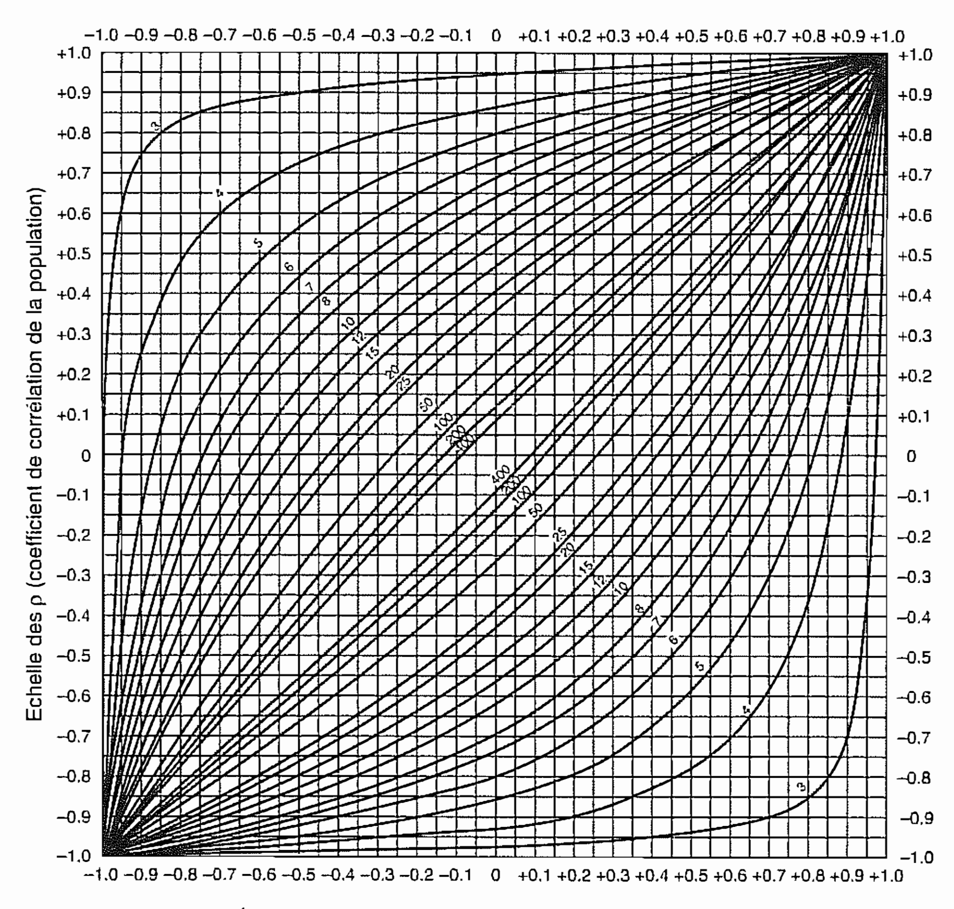
ν \ α	0,1	0,05	0,01	0,001	ν \ α	0,1	0,05	0,01	0,001
101	0,1630	0,1937	0,2528	0,3196	128	1449	1723	2252	2854
102	1622	1927	2515	3181	129	1443	1716	2243	2843
103	1614	1918	2504	3166	**130**	0,1438	0,1710	0,2235	0,2832
104	1606	1909	2492	3152	131	1432	1703	2226	2822
105	0,1599	0,1900	0,2480	0,3138	132	1427	1697	2218	2812
106	1591	1891	2469	3123	133	1422	1690	2210	2801
107	1584	1882	2458	3109	134	1416	1684	2202	2791
108	1577	1874	2447	3095	**135**	0,1411	0,1678	0,2194	0,2781
109	1569	1865	2436	3082	136	1406	1672	2186	2771
110	0,1562	0,1857	0,2425	0,3069	137	1401	1666	2178	2762
111	1555	1848	2414	3055	138	1396	1660	2170	2752
112	1548	1840	2404	3042	139	1391	1654	2163	2742
113	1542	1832	2393	3029	**140**	0,1386	0,1648	0,2155	0,2733
114	1535	1824	2383	3017	141	1381	1642	2148	2724
115	0,1528	0,1816	0,2373	0,3004	142	1376	1637	2140	2714
116	1522	1809	2363	2992	143	1371	1631	2133	2705
117	1515	1801	2353	2979	144	1367	1625	2126	2696
118	1509	1793	2343	2967	**145**	0,1362	0,1620	0,2118	0,2687
119	1502	1786	2334	2955	146	1357	1614	2111	2678
120	0,1496	0,1779	0,2324	0,2943	147	1353	1609	2104	2669
121	1490	1771	2315	2932	148	1348	1603	2097	2660
122	1484	1764	2305	2920	149	1344	1598	2090	2652
123	1478	1757	2296	2909	**150**	0,1339	0,1593	0,2083	0,2643
124	1472	1750	2287	2897	151	1335	1587	2077	2635
125	0,1466	0,1743	0,2278	0,2886	152	1330	1582	2070	2626
126	1460	1736	2269	2875	153	1326	1577	2063	2618
127	1455	1730	2261	2864	154	1322	1572	2057	2610

TABLE A.9 (*suite et fin*) VALEURS CRITIQUES DU COEFFICIENT DE CORRÉLATION R D'UN ÉCHANTILLON ISSU D'UNE POPULATION NORMALE OÙ $\rho = 0$
Cette table donne les valeurs r telles que $P(|R| > r) = \alpha$, $\nu = n - 2$ corrélation simple
$\nu = n - 2 - d$ corrélation partielle avec d variables fixées

ν \ α	0,1	0,05	0,01	0,001	ν \ α	0,1	0,05	0,01	0,001
155	0,1318	0,1567	0,2050	0,2602	178	1230	1463	1915	2433
156	1313	1562	2044	2594	179	1227	1459	1910	2426
157	1309	1557	2037	2586	**180**	0,1223	0,1455	0,1905	0,2420
158	1305	1552	2031	2578	181	1220	1451	1900	2413
159	1301	1547	2025	2570	182	1216	1447	1895	2407
160	0,1297	0,1543	0,2019	0,2562	183	1213	1443	1890	2400
161	1293	1538	2012	2554	184	1210	1439	1885	2394
162	1289	1533	2006	2547	**185**	0,1207	0,1435	0,1880	0,2388
163	1285	1529	2000	2539	186	1203	1432	1874	2381
164	1281	1524	1994	2532	187	1200	1428	1870	2375
165	0,1277	0,1519	0,1988	0,2524	188	1197	1424	1865	2369
166	1273	1515	1982	2517	189	1194	1420	1860	2363
167	1270	1510	1977	2510	**190**	0,1191	0,1417	0,1855	0,2357
168	1266	1506	1971	2502	191	1188	1413	1850	2351
169	1262	1501	1965	2495	192	1184	1409	1845	2345
170	0,1258	0,1497	0,1959	0,2488	193	1181	1406	1841	2339
171	1255	1493	1954	2481	194	1178	1402	1836	2333
172	1251	1488	1948	2474	**195**	0,1175	0,1399	0,1831	0,2327
173	1248	1484	1943	2467	196	1172	1395	1827	2321
174	1244	1480	1937	2460	197	1169	1391	1822	2316
175	0,1240	0,1476	0,1932	0,2453	198	1166	1388	1818	2310
176	1237	1471	1926	2446	199	1164	1384	1813	2304
177	1233	1467	1921	2440	**200**	0,1161	0,1381	0,1809	0,2299

Pour $\nu > 200$ on admet que r est une réalisation d'une variable de Laplace-Gauss d'espérance nulle et d'écart-type $\frac{1}{\sqrt{\nu + 1}}$.

TABLE A.9 *bis* INTERVALLES DE CONFIANCE POUR LE COEFFICIENT DE CORRÉLATION
(Niveau de confiance .95)



Échelle des r (coefficient de corrélation de l'échantillon)
Les nombres sur les courbes correspondent aux tailles d'échantillon

TABLE A.10 TABLE DE CORRESPONDANCE ENTRE r ET z

(Corrélation transformée de R. A. Fisher)

$$r = \frac{\exp(2x) - 1}{\exp(2x) + 1}$$

$$z = \frac{1}{2} \ln \frac{1 + r}{1 - r}$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0.0300	0,0400	0,0500	0,0599	0,0699	0,0798	0,0898
0,1	0,0997	0,1096	0,1194	0,1293	0.1391	0,1489	0,1586	0,1684	0,1781	0,1877
0,2	0,1974	0,2070	0,2165	0.2260	0,2355	0,2449	0,2548	0,2636	0,2729	0.2821
0,3	0,2913	0,3004	0,3095	0.3185	0,3275	0,3364	0,3452	0,3540	0.3627	0.3714
0.4	0,3800	0,3885	0.3969	0,4053	0.4136	0,4219	0.4301	0,4382	0,4462	0,4542
0,5	0,4621	0,4699	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299
0,6	0,5370	0,5441	0.5511	0,5580	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980
0,7	0,6044	0,6107	0,6169	0,6231	0,6291	0.6351	0.6411	0,6469	0,6527	0,6584
0,8	0,6640	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	0,7114
0,9	0,7163	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574
1,0	0.7616	0.7658	0,7699	0,7739	0,7779	0,7818	0,7857	0.7895	0,7932	0,7969
1.1	0,8005	0,8041	0,8076	0,8110	0,8144	0,8178	0,8210	0,8243	0,8275	0,8306
1,2	0,8337	0,8367	0,8397	0,8426	0.8455	0,8483	0,8511	0,8538	0,8565	0,8591
1,3	0,8617	0,8643	0,8668	0,8692	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832
1,4	0,8854	0.8875	0,8896	0,8917	0.8937	0,8957	0,8977	0.8996	0,9015	0,9033
1,5	0,9051	0.9069	0,9087	0,9104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9201
1,6	0,9217	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9341
1,7	0,9354	0,9336	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458
1,8	0.9468	0,94783	0,94884	0,94983	0,95080	0,95175	0,95268	0.95359	0,95449	0,95537
1,9	0,95624	0,95709	0,95792	0,95873	0,95953	0,96032	0,96109	0,96185	0,96259	0,96331
2,0	0,96403	0,96473	0,96541	0,96609	0,96675	0,96739	0,96803	0,96865	0,96926	0,96986
2,1	0,97045	0,97103	0,97159	0,97215	0,97269	0,97323	0,97375	0,97426	0,97477	0,97526
2,2	0,97574	0,97622	0,97668	0,97714	0,97752	0,97803	0,97846	0,97888	0,97929	0,97970
2,3	0,98010	0,98049	0,98087	0,98124	0,98161	0.98197	0,98233	0,98267	0,98301	0,98335
2,4	0,98367	0,98399	0,98431	0,98462	0,98492	0,98522	0,98551	0,98579	0,98607	0.98635
2,5	0.98661	0,98688	0,98714	0,98739	0,98764	0,98788	0,98812	0,98835	0,98858	0,98881
2,6	0,98903	0,98924	0,98945	0,98966	0.98987	0,99007	0,99026	0,99045	0,99064	0,99083
2,7	0,99101	0,99118	0,99136	0,99153	0.99170	0,99185	0,99202	0,99218	0,99233	0,99248
2,8	0,99263	0,99278	0,99292	0.99306	0,99320	0,99333	0,99346	0,99359	0,99372	0,99384
2,9	0,99396	0,99408	0,99420	0,99431	0,99443	0,99454	0,99464	0,99475	0,99485	0,99495

TABLE A.11 TABLE DU COEFFICIENT DE CORRÉLATION DES RANGS DE SPEARMAN ENTRE DE DEUX VARIABLES INDÉPENDANTES

Valeurs r de R_s ayant une probabilité α d'être dépassée en valeur absolue

$$P(|R_s| > r) = \alpha$$

n \ α	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
4	0.600	1.000	1.000						
5	0.500	0.800	0.900	1.000	1.000				
6	0.371	0.657	0.829	0.886	0.943	1.000	1.000		
7	0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000
8	0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976
9	0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933
10	0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903
11	0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873
12	0.224	0.406	0.503	0.587	0.671	0.727	0.776	0.825	0.860
13	0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.802	0.835
14	0.200	0.367	0.464	0.538	0.622	0.675	0.723	0.776	0.811
15	0.189	0.354	0.443	0.521	0.604	0.654	0.700	0.754	0.786
16	0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.732	0.765
17	0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748
18	0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728
19	0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712
20	0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696
21	0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681
22	0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667
23	0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654
24	0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642
25	0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630
26	0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619
27	0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608
28	0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598
29	0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589
30	0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580
31	0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571
32	0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563
33	0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554
34	0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547
35	0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539
36	0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533
37	0.114	0.216	0.275	0.325	0.383	0.421	0.456	0.497	0.526
38	0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519
39	0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513
40	0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507

Table A.11 *(suite et fin)* Table du coefficient de corrélation des rangs de Spearman de deux variables indépendantes

Valeurs r de R_s ayant une probabilité α d'être dépassée en valeur absolue

$P(|R_s| > r) = \alpha$

n \ α	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
41	0.108	0.204	0.261	0.309	0.364	0.400	0.433	0.473	0.501
42	0.107	0.202	0.257	0.305	0.359	0.395	0.428	0.468	0.495
43	0.105	0.199	0.254	0.301	0.355	0.391	0.423	0.463	0.490
44	0.104	0.197	0.251	0.298	0.351	0.386	0.419	0.458	0.484
45	0.103	0.194	0.248	0.294	0.347	0.382	0.414	0.453	0.479
46	0.102	0.192	0.246	0.291	0.343	0.378	0.410	0.448	0.474
47	0.101	0.190	0.243	0.288	0.340	0.374	0.405	0.443	0.469
48	0.100	0.188	0.240	0.285	0.336	0.370	0.401	0.439	0.465
49	0.098	0.186	0.238	0.282	0.333	0.366	0.397	0.434	0.460
50	0.097	0.184	0.235	0.279	0.329	0.363	0.393	0.430	0.456
52	0.095	0.180	0.231	0.274	0.323	0.356	0.386	0.422	0.447
54	0.094	0.177	0.226	0.268	0.317	0.349	0.379	0.414	0.439
56	0.092	0.174	0.222	0.264	0.311	0.343	0.372	0.407	0.432
58	0.090	0.171	0.218	0.259	0.306	0.337	0.366	0.400	0.424
60	0.089	0.168	0.214	0.255	0.300	0.331	0.360	0.394	0.418
62	0.087	0.165	0.211	0.250	0.296	0.326	0.354	0.388	0.411
64	0.086	0.162	0.207	0.246	0.291	0.321	0.348	0.382	0.405
66	0.084	0.160	0.204	0.243	0.287	0.316	0.343	0.376	0.399
68	0.083	0.157	0.201	0.239	0.282	0.311	0.338	0.370	0.393
70	0.082	0.155	0.198	0.235	0.278	0.307	0.333	0.365	0.388
72	0.081	0.153	0.195	0.232	0.274	0.303	0.329	0.360	0.382
74	0.080	0.151	0.193	0.229	0.271	0.299	0.324	0.355	0.377
76	0.078	0.149	0.190	0.226	0.267	0.295	0.320	0.351	0.372
78	0.077	0.147	0.188	0.223	0.264	0.291	0.316	0.346	0.368
80	0.076	0.145	0.185	0.220	0.260	0.287	0.312	0.342	0.363
82	0.075	0.143	0.183	0.217	0.257	0.284	0.308	0.338	0.359
84	0.074	0.141	0.181	0.215	0.254	0.280	0.305	0.334	0.355
86	0.074	0.139	0.179	0.212	0.251	0.277	0.301	0.330	0.351
88	0.073	0.138	0.176	0.210	0.248	0.274	0.298	0.327	0.347
90	0.072	0.136	0.174	0.207	0.245	0.271	0.294	0.323	0.343
92	0.071	0.135	0.173	0.205	0.243	0.268	0.291	0.319	0.339
94	0.070	0.133	0.171	0.203	0.240	0.265	0.288	0.316	0.336
96	0.070	0.132	0.169	0.201	0.238	0.262	0.285	0.313	0.332
98	0.069	0.130	0.167	0.199	0.235	0.260	0.282	0.310	0.329
100	0.068	0.129	0.165	0.197	0.233	0.257	0.279	0.307	0.326

Pour $n > 100$ on admet que R_s est distribué comme $LG\left(0\,;\frac{1}{\sqrt{n-1}}\right)$.

TABLE A.12 TEST DE CONCORDANCE DE *p* CLASSEMENTS
(test du W de M. G. Kendall)
Valeurs critiques w de W à $\alpha = 0.05$
$P(W \geq w) = 0.05$

n \ p	3	4	5	6
3	1	0,750	0,600	0,500
4	0,822	0,619	0,500	0,421
5	0,716	0,553	0,449	0,377
6	0,660	0,512	0,418	0,351
7	0,626	0,484	0,395	0,332
8	0,595	0,461	0,378	0,319
9	0,576	0,447	0,365	0,307
10	0,560	0.434	0,354	0,299
11	0,548	0,425	0,346	0,287
12	0,535	0,415	0,336	0,287
13	0,527	0,409	0,332	0,280
14	0,520	0,402	0,327	0,275
15	0,514	0,395	0,322	0,272
20	0,49	0,37	0,30	0,25
40	0,43	0,33	0,26	0,22
60	0,41	0,31	0,25	0,21
100	0,38	0,29	0,24	0,20
∞	0,33	0,25	0,20	0,17

Pour $p \geq 7$ la quantité $p(n-1)W$ est distribuée approximativement selon un χ^2_{n-1}.

TABLE A.13 FONCTION DE RÉPARTITION DE LA STATISTIQUE DE CRAMER-VON MISES

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(\frac{2i-1}{2n} - F(y_i)\right)^2 = \int_{-\infty}^{+\infty} (F_n^*(x) - F(x))^2 \, \mathrm{d}F(x)$$

F_n^* est la fonction de répartition empirique de l'échantillon

F est la fonction de répartition de la variable échantillonnée $y_1, y_2, \ldots, y_n$ les valeurs de l'échantillon ordonné

Cette table donne les valeurs z telles que : $1 - \alpha = P(n\omega_n^2 < z)$.

	$1-\alpha$														
n	0.99	0.975	0.95	0.90	0.85	0.80	0.75	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01
2	0.55052	0.48897	0.42482	0.34346	0.28853	0.24743	0.21521	0.12659	0.08145	0.07351	0.06554	0.05758	0.04963	0.04565	0.04326
3	0.63976	0.53316	0.43938	0.33786	0.27963	0.24169	0.21339	0.12542	0.07683	0.06886	0.06092	0.05287	0.04355	0.03777	0.03324
4	0.67017	0.54200	0.44199	0.34183	0.28337	0.24260	0.21173	0.12405	0.07494	0.06681	0.05895	0.05093	0.04147	0.03537	0.03013
5	0.68352	0.55056	0.44697	0.34238	0.28305	0.24236	0.21165	0.12252	0.07427	0.06611	0.05799	0.04970	0.04035	0.03422	0.02876
6	0.69443	0.55572	0.44911	0.34352	0.28331	0.24198	0.21110	0.12200	0.07352	0.06548	0.05747	0.04910	0.03960	0.03344	0.02794
7	0.70154	0.55935	0.45100	0.34397	0.28345	0.24197	0.21087	0.12158	0.07297	0.06492	0.05697	0.04869	0.03914	0.03293	0.02738
8	0.70912	0.56327	0.45285	0.34462	0.28358	0.24187	0.21066	0.12113	0.07254	0.06448	0.05650	0.04823	0.03876	0.03256	0.02706
9	0.71283	0.56513	0.45377	0.34491	0.28364	0.24180	0.21052	0.12088	0.07228	0.06423	0.05625	0.04798	0.03850	0.03230	0.02679
10	0.71582	0.56663	0.45450	0.34514	0.28368	0.24175	0.21041	0.12069	0.07208	0.06403	0.05605	0.04778	0.03830	0.03209	0.02657
20	0.72948	0.57352	0.45788	0.34621	0.28387	0.24150	0.20990	0.11979	0.07117	0.06312	0.05515	0.04689	0.03742	0.03120	0.02564
50	0.73784	0.57775	0.45996	0.34686	0.28398	0.24134	0.20960	0.11924	0.07062	0.06258	0.05462	0.04636	0.03690	0.03068	0.02512
200	0.74205	0.57990	0.46101	0.34719	0.28404	0.24126	0.20944	0.11897	0.07035	0.06231	0.05435	0.04610	0.03665	0.03043	0.02488
1000	0.74318	0.58047	0.46129	0.34728	0.28406	0.24124	0.20940	0.11890	0.07027	0.06224	0.05428	0.04603	0.03658	0.03037	0.02481
∞	0.74346	0.58061	0.46136	0.34730	0.28406	0.24124	0.20939	0.11888	0.07026	0.06222	0.05426	0.04601	0.03656	0.03035	0.02480

TABLE A.14 TABLE DU TEST DE KOLMOGOROV-SMIRNOV

$D_n = \sup |F_n^*(x) - F(x)|$

Valeurs de d_n telles que $P = P(D_n < d_n)$

n	$P = .80$	$P = .90$	$P = .95$	$P = .98$	$P = .99$
1	.90000	.95000	.97500	.99000	.99500
2	.68377	.77639	.84189	.90000	.92929
3	.56481	.63604	.70760	.78456	.82900
4	.49265	.56522	.62394	.68887	.73424
5	.44698	.50945	.56328	.62718	.66853
6	.41037	.46799	.51926	.57741	.61661
7	.38148	.43607	.48342	.53844	.57581
8	.35831	.40962	.45427	.50654	.54179
9	.33910	.38746	.43001	.47960	.51332
10	.32260	.36866	.40925	.45662	.48893
11	.30829	.35242	.39122	.43670	.46770
12	.29577	.33815	.37543	.41918	.44905
13	.28470	.32549	.36143	.40362	.43247
14	.27481	.31417	.34890	.38970	.41762
15	.26588	.30397	.33760	.37713	.40420
16	.25778	.29472	.32733	.36571	.39201
17	.25039	.28627	.31796	.35528	.38086
18	.24360	.27851	.30936	.34569	.37062
19	.23735	.27136	.30143	.33685	.36117
20	.23156	.26473	.29408	.32866	.35241
21	.22617	.25858	.28724	.32104	.34427
22	.22115	.25283	.28087	.31394	.33666
23	.21645	.24746	.27490	.30728	.32954
24	.21205	.24242	.26931	.30104	.32286
25	.20790	.23768	.26404	.29516	.31657
26	.20399	.23320	.25907	.28962	.31064
27	.20030	.22898	.25438	.28438	.30502
28	.19680	.22497	.24993	.27942	.29971
29	.19348	.22117	.24571	.27471	.29466
30	.19032	.21756	.24170	.27023	.28987
31	.18732	.21412	.23788	.26596	.28530
32	.18445	.21085	.23424	.26189	.28094
33	.18171	.20771	.23076	.25801	.27677
34	.17909	.20472	.22743	.25429	.27279
35	.17659	.20185	.22425	.25073	.26897
36	.17418	.19910	.22119	.24732	.26532
37	.17188	.19646	.21826	.24404	.26180
38	.16966	.19392	.21544	.24089	.25843
39	.16753	.19148	.21273	.23786	.25518
40	.16547	.18913	.21012	.23494	.25205
41	.16349	.18687	.20760	.23213	.24904
42	.16158	.18468	.20517	.22941	.24613
43	.15974	.18257	.20283	.22679	.24332
44	.15796	.18053	.20056	.22426	.24060
45	.15623	.17856	.19837	.22181	.23798
46	.15457	.17665	.19625	.21944	.23544
47	.15295	.17481	.19420	.21715	.23298
48	.15139	.17302	.19221	.21493	.23059
49	.14987	.17128	.19028	.21277	.22828
50	.14840	.16959	.18841	.21068	.22604

TABLE A.14 *(suite et fin)* TABLE DU TEST DE KOLMOGOROV-SMIRNOV

$D_n = \sup |F_n^*(x) - F(x)|$

Valeurs de d_n telles que $P = P(D_n < d_n)$

n	$P = .80$	$P = .90$	$P = .95$	$P = .98$	$P = .99$
51	.14697	.16796	.18659	.20864	.22386
52	.14558	.16637	.18482	.20667	.22174
53	.14423	.16483	.18311	.20475	.21968
54	.14292	.16332	.18144	.20289	.21768
55	.14164	.16186	.17981	.20107	.21574
56	.14040	.16044	.17823	.19930	.21384
57	.13919	.15906	.17669	.19758	.21199
58	.13801	.15771	.17519	.19590	.21019
59	.13686	.15639	.17373	.19427	.20844
60	.13573	.15511	.17231	.19267	.20673
61	.13464	.15385	.17091	.19112	.20506
62	.13357	.15263	.16956	.18960	.20343
63	.13253	.15144	.16823	.18812	.20184
64	.13151	.15027	.16693	.18667	.20029
65	.13052	.14913	.16567	.18525	.19877
66	.12954	.14802	.16443	.18387	.19729
67	.12859	.14693	.16322	.18252	.19584
68	.12766	.14587	.16204	.18119	.19442
69	.12675	.14483	.16088	.17990	.19303
70	.12586	.14381	.15975	.17863	.19167
71	.12499	.14281	.15864	.17739	.19034
72	.12413	.14183	.15755	.17618	.18903
73	.12329	.14087	.15649	.17498	.18776
74	.12247	.13993	.15544	.17382	.18650
75	.12167	.13901	.15442	.17268	.18528
76	.12088	.13811	.15342	.17155	.18408
77	.12011	.13723	.15244	.17045	.18290
78	.11935	.13636	.15147	.16938	.18174
79	.11860	.13551	.15052	.16832	.18060
80	.11787	.13467	.14960	.16728	.17949
81	.11716	.13385	.14868	.16626	.17840
82	.11645	.13305	.14779	.16526	.17732
83	.11576	.13226	.14691	.16428	.17627
84	.11508	.13148	.14605	.16331	.17523
85	.11442	.13072	.14520	.16236	.17421
86	.11376	.12997	.14437	.16143	.17321
87	.11311	.12923	.14355	.16051	.17223
88	.11248	.12850	.14274	.15961	.17126
89	.11186	.12779	.14195	.15873	.17031
90	.11125	.12709	.14117	.15786	.16938
91	.11064	.12640	.14040	.15700	.16846
92	.11005	.12572	.13965	.15616	.16755
93	.10947	.12506	.13891	.15533	.16666
94	.10889	.12440	.13818	.15451	.16579
95	.10833	.12375	.13746	.15371	.16493
96	.10777	.12312	.13675	.15291	.16408
97	.10722	.12249	.13606	.15214	.16324
98	.10668	.12187	.13537	.15137	.16242
99	.10615	.12126	.13469	.15061	.16161
100	.10563	.12067	.13403	.14987	.16081
n > 100	$1.073/\sqrt{n}$	$1.223/\sqrt{n}$	$1.358/\sqrt{n}$	$1.518/\sqrt{n}$	$1.629/\sqrt{n}$

TABLE A.15 VALEURS CRITIQUES DU COEFFICIENT D'ASYMÉTRIE EMPIRIQUE D'UN ÉCHANTILLON DE n OBSERVATIONS D'UNE VARIABLE DE LAPLACE-GAUSS

$$P\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3} > b\right) = \alpha$$

n \ α	0.05	0.01
7	1.018	1.457
8	0.998	1.452
9	0.977	1.433
10	0.954	1.407
12	0.910	1.353
15	0.851	1.272
20	0.772	1.155
25	0.711	1.061
30	0.662	0.986
35	0.621	0.923
40	0.587	0.870
45	0.558	0.825
50	0.534	0.787
60	0.492	0.723
70	0.459	0.673
80	0.432	0.631
90	0.409	0.596
100	0.389	0.567
125	0.350	0.508
150	0.321	0.464
175	0.298	0.430
200	0.280	0.403
250	0.251	0.360
300	0.230	0.329

n \ α	0.05	0.01
350	0.213	0.305
400	0.200	0.285
450	0.188	0.269
500	0.179	0.255
550	0.171	0.243
600	0.163	0.233
650	0.157	0.224
700	0.151	0.215
750	0.146	0.208
800	0.142	0.202
850	0.138	0.196
900	0.134	0.190
950	0.130	0.185
1000	0.127	0.180
1200	0.116	0.165
1400	0.107	0.152
1600	0.100	0.142
1800	0.095	0.134
2000	0.090	0.127
2500	0.080	0.114
3000	0.073	0.104
3500	0.068	0.096
4000	0.064	0.090
4500	0.060	0.085
5000	0.057	0.081

TABLE A.16 VALEURS CRITIQUES DE COEFFICIENT D'APLATISSEMENT D'UN ÉCHANTILLON DE n OBSERVATIONS D'UNE VARIABLE DE LAPLACE-GAUSS

$$P\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{s^4} < b\right) = \alpha$$

n \ α	0.01	0.05	0.95	0.99
7	1.25	1.41	3.55	4.29
8	1.31	1.46	3.70	4.53
9	1.35	1.53	3.86	4.82
10	1.34	1.56	3.95	5.00
12	1.46	1.64	4.05	5.20
15	1.55	1.72	4.13	5.30
20	1.65	1.82	4.17	5.36
25	1.72	1.91	4.16	5.30
30	1.79	1.98	4.11	5.21
35	1.84	2.03	4.10	5.13
40	1.89	2.07	4.06	5.04
45	1.93	2.11	4.00	4.94
50	1.95	2.15	3.99	4.88
75	2.08	2.27	3.87	4.59
100	2.18	2.35	3.77	4.39
125	2.24	2.40	3.71	4.24
150	2.29	2.45	3.65	4.13
200	2.37	3.51	3.57	3.98
250	2.42	2.55	3.52	3.87
300	2.46	2.59	3.47	3.79
350	2.50	2.62	3.44	3.72
400	2.52	2.64	3.41	3.67
450	2.55	2.66	3.39	3.63
500	2.57	2.67	3.37	3.60
550	2.58	2.69	3.35	3.57
600	2.60	2.70	3.34	3.54
650	2.61	2.71	3.33	3.52
700	2.62	2.72	3.31	3.50
800	2.65	2.74	3.29	3.46
900	2.66	2.75	3.28	3.43
1000	2.68	2.76	3.26	3.41
1200	2.71	2.78	3.24	3.37
1400	2.72	2.80	3.22	3.34
1600	2.74	2.81	3.21	3.32
1800	2.76	2.82	3.20	3.30
2000	2.77	2.83	3.18	3.28
2500	2.79	2.85	3.16	3.25
3000	2.81	2.86	3.15	3.22
3500	2.82	2.87	3.14	3.21
4000	2.83	2.88	3.13	3.19
4500	2.84	2.88	3.12	3.18
5000	2.85	2.89	3.12	3.17

TABLE A.17 TEST DE DURBIN ET WATSON
VALEURS CRITIQUES AU SEUIL 5 % POUR $H_0 : \rho = 0$
p : nombre de variables explicatives
n : nombre d'observations

d_{inf} d_{sup}

0 H_0 refusée | incertitude | H_0 acceptée 2

n	$p = 1$		$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	d_{inf}	d_{sup}	d_{inf}	d_{sup}	d_{inf}	d_{sup}	d_{inf}	d_{sup}	d_{inf}	d_{sup}
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

TABLE A.18 COEFFICIENTS POUR CALCULER L'ESPÉRANCE ET LA VARIANCE DE L'ÉCART-TYPE CORRIGÉ ET DE L'ÉTENDUE D'UN ÉCHANTILLON GAUSSIEN

n	c_4	d_2	d_3
2	0.7979	1.128	0.853
3	0.8862	1.693	0.888
4	0.9213	2.059	0.880
5	0.9400	2.326	0.864
6	0.9515	2.534	0.848
7	0.9594	2.704	0.833
8	0.9650	2.847	0.820
9	0.9693	2.970	0.808
10	0.9727	3.078	0.797
11	0.9754	3.173	0.787
12	0.9776	3.258	0.778
13	0.9794	3.336	0.770
14	0.9810	3.407	0.762
15	0.9823	3.472	0.755

$E(S^*) = c_4\sigma \qquad V(S^*) = (1 - c_4^2)\sigma^2$

$E(R) = d_2\sigma \qquad V(R) = (d_3\sigma)^2$

Voir chapitre 12 § 12.2.3.3.

B Formulaire

Tableau B.1 Paramètres des principales distributions discrètes

Loi	Espérance $E(X)$	Variance $V(X)$	Coefficient d'asymétrie γ_1	Coefficient d'aplatissement γ_2
Binomiale $\mathscr{B}(n\,;p)$ $P(X=x) = C_n^x p^x q^{n-x}$ $X = 0, 1, 2, \ldots, n$	np	npq	$\frac{q-p}{\sqrt{npq}}$	$3 + \frac{1-6pq}{npq}$
Binomiale négative $B^-(n\,;p)$ $P(X=x) = C_{n+x-1}^{n-1}\left(\frac{p}{q}\right)^x\left(1-\frac{p}{q}\right)^n$ $= C_{-n}^x p^x q^{-n-x}$ $q = 1 + p$ $X = 0, 1, 2, \ldots, \infty$	np	npq	$\frac{p+q}{\sqrt{npq}}$	$3 + \frac{1+6pq}{npq}$
Pascal Pa$(n\,;p)$ $P(X=x) = C_{x-1}^{n-1} p^n q^{x-n}$ $X = n, n+1, \ldots, \infty \quad p+q=1$	$\frac{n}{p}$	$\frac{nq}{p^2}$	$\frac{2-p}{\sqrt{nq}}$	$3 + \frac{p^2+6q}{nq}$
Hypergéométrique $\mathscr{H}(N, n, p)$ $P(X=x) = \frac{C_{Np}^x C_{Nq}^{n-x}}{C_N^n}$	np	$npq\frac{N-n}{N-1}$	$\frac{q-p}{\sqrt{npq}} \cdot \frac{N-2n}{N-2}\sqrt{\frac{N-1}{N-n}}$	$3\frac{(N-1)(N+6)}{(N-2)(N-3)} + \frac{(N-1)N(N+1)}{(N-n)(N-2)(N-3)} \cdot \frac{1}{npq}\left(1-6\frac{N}{N+1}\left(pq+\frac{n(n-N)}{N^2}\right)\right)$
Poisson $\mathscr{P}(m)$ $P(X=x) = \exp(-m)\left(\frac{m^x}{x!}\right)$ $X = 0, 1, 2, \ldots, \infty$	m	m	$\frac{1}{\sqrt{m}}$	$3 + \frac{1}{m}$
Uniforme $P(X=x) = \frac{1}{n}$ $X = 1, 2, \ldots, n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	0	$1.8 - \frac{2.4}{n^2-1}$

TABLEAU B.2 PARAMÈTRES DES PRINCIPALES DISTRIBUTIONS CONTINUES

Loi	Espérance $E(X)$	Variance $V(X)$	Coefficient d'asymétrie $\gamma_1 = \frac{\mu_3}{\sigma^3}$	Coefficient d'aplatissement $\gamma_2 = \frac{\mu_4}{\sigma^4}$
Continue uniforme sur [0, 1]	$1/2$	$1/12$	0	1.8
$LG(m\ ;\ \sigma)$	m	σ^2	0	3
γ_r	r	r	$2/\sqrt{r}$	$3 + 6/r$
χ^2_n	n	$2n$	$\sqrt{8/n}$	$3 + 12/n$
Student T_n	0	$n/(n-2)$ (1)	0	$3 + 6/(n-4)$ (2)
Beta I (n, p)	$n/(n+p)$	$\frac{np}{(n+p+1)(n+p)}$	$\frac{2(p-n)\sqrt{n^{-1}+p^{-1}+(np)^{-1}}}{n+p+2}$	$\frac{(3(n+p+1)(2(n+p)^2+np(n+p-6))}{np(n+p+2)(n+p+3)}$
Beta II (n, p)	$\frac{n}{p-1}$	$\frac{n(n+p-1)}{(p-1)^2(p-2)}$	$2\sqrt{\frac{(p-2)}{n(n+p-1)}}\frac{2n+p-1}{p-3}$	$\frac{6(p-1)^2(p-2)+n(n+p-1)(5p-11)}{n(p-3)(p-4)(n+p-1)} + 3$
$F(n, p)$	$\frac{n}{p-2}$	$\frac{2p^2(n+p-2)}{n(p-2)^2(p-4)}$	$\sqrt{\frac{8(p-4)}{n(n+p-2)}}\frac{2n+p-2}{p-6}$	$\frac{12(p-2)^2(p-4)+m(n+p-2)(5p-22)}{n(p-6)(p-8)(n+p-2)} + 3$
Log-normale	$\exp\left(m + \frac{\sigma^2}{2}\right)$	$\exp(2m+\sigma^2)(\exp\sigma^2 - 1)$	$(\exp\sigma^2 + 2)\sqrt{\exp\sigma^2 - 1}$	$\exp 4\sigma^2 + 2\exp 3\sigma^2 + 3\exp 2\sigma^2 - 3$
Weibull $f(x) = \beta x^{\beta-1}\exp(-x^{\beta})$	$\Gamma\left(1+\frac{1}{\beta}\right)$	$\Gamma\left(1+\frac{2}{\beta}\right) - (E(x))^2$		
Gumbel $\exp(-x - \exp(-x))$	0.57722	$\pi^2/6$	1.29857	5.4

(1) si $n > 2$.
(2) si $n > 4$.

Quelques relations exactes entre les principales distributions

Loi de Pascal et loi binomiale négative

Si X suit une loi Pa $(n ; p)$, $X - n$ suit une loi binomiale négative $B^{-}\left(n ; \frac{1-p}{p}\right)$.

Loi de Poisson et loi du χ^2

Si X suit une loi $\mathcal{P}(m)$: $\quad P(X \leq x) = P(\chi^2_{2(x+1)} > 2m)$

Loi binomiale et loi de Fisher-Snedecor

Si X suit une loi $\mathcal{B}(n ; p)$: $\quad P(X \leq x) = P\left(F > \frac{n-x}{x+1}\frac{p}{1-p}\right)$

où F a pour degré de liberté $2(x + 1)$ et $2(n - x)$.

Loi de Fisher-Snedecor et loi de Student

$T_n^2 = F(1 ; n)$

Loi gamma et loi du χ^2

Si X suit une loi γ_r, $2X$ est un χ^2_{2r}.

Lois bêta et loi de Fisher-Snedecor

Si X bêta I $(n ; p)$: $\quad \frac{p}{n}\frac{X}{1-X} = F(2n ; 2p)$

Si X bêta II $(n ; p)$: $\quad \frac{pX}{n} = F(2n ; 2p)$



Calcul des fonctions de répartition de certaines lois continues

Les formules qui suivent permettent de calculer exactement ou approximativement avec une grande précision $P(X < x)$ ou $P(X > x)$. Leur intérêt est d'être facilement programmables même sur une calculatrice de poche et d'éviter le recours à des tables.

C.1 LOI NORMALE CENTRÉE-RÉDUITE

L'approximation suivante fournit pour tout u positif $P(U < u)$ avec une erreur inférieure à 10^{-7}.

$$P(U < u) = 1 - f(u)\,(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5)$$

avec :

$$f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) ;$$

$$t = \frac{1}{1 + 0.2316419u} ;$$

$$b_1 = 0.319381530 ;$$

$$b_2 = -0.356563782 ;$$

$$b_3 = 1.781477937 ;$$

$$b_4 = -1.821255978 ;$$

$$b_5 = 1.330274429.$$

C.2 LOI DU χ^2_ν

C.2.1 Formules exactes

C.2.1.1 ν *pair :*

$$P(\chi^2_\nu > x) = \sum_{i=0}^{\nu/2-1} \exp\left(-\frac{x}{2}\right)\frac{\left(\frac{x}{2}\right)^i}{i!}$$

en particulier on a $P(\chi^2_2 < x) = 1 - \exp\left(-\frac{x}{2}\right)$.

C.2.1.2 ν *impair :*

$$P(\chi^2_\nu > x) = 2P\left(U > \sqrt{x}\right) + \frac{2}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x\right)\sum_{j=1}^{\frac{\nu-1}{2}} \frac{x^{j-\frac{1}{2}}}{1.3.5\ldots(2j-1)}$$

en particulier on a $P(\chi^2_1 < x) = 2P\left(U < \sqrt{x}\right) - 1$.

C.2.2 Formules approchées

La meilleure est celle de Wilson-Hilferty qui donne P avec deux décimales exactes dès que $n \geq 3$:

$$P(\chi^2_\nu < x) \simeq P\left(U < \left(\frac{9\nu}{2}\right)^{\frac{1}{2}}\left(\left(\frac{x}{\nu}\right)^{\frac{1}{3}} + \frac{2}{9\nu} - 1\right)\right)$$

d'où pour la valeur x de χ^2_ν ayant la probabilité p d'être dépassée :

$$x \simeq \nu\left[1 - \frac{2}{9\nu} + u\sqrt{\frac{2}{9\nu}}\right]^3$$

où u est le fractile correspondant de U : $P(U > u) = p$.

C.3 LOI DE FISHER-SNEDECOR $F(\nu_1\,;\,\nu_2)$

On utilisera l'approximation de Paulson dérivée de celle de Wilson-Hilferty :

$$P(F < f) \simeq P\left(U < \frac{f^{1/3}\left(1 - \frac{2}{9\nu_2}\right) + \frac{2}{9\nu_1} - 1}{\sqrt{\frac{2}{9\nu_1} + f^{2/3}\frac{2}{9\nu_2}}}\right)$$

elle donne dès que $\nu_2 \geq 4$ et pour tout ν_1, 2 décimales exactes.

Si $\nu_2 \leq 3$ on multipliera la fraction ci-dessus par $\left(1 + 0.08 \frac{u^4}{\nu_2^3}\right)$.

C.4 FONCTION DE RÉPARTITION DE LA LOI DE STUDENT T_n

C.4.1 Formules exactes

C.4.1.1 $n = 1$

T_1 est la loi de Cauchy de densité $\frac{1}{\pi(1 + t^2)}$ d'où :

$$P(T < t) = \frac{1}{2} + \frac{1}{\pi} \text{Arc tg } t$$

réciproquement si on connaît α tel que : $P(|T_1| < t)$ on a : $t = \text{tg}\left(\frac{\pi}{2}\alpha\right)$.

C.4.1.2 $n \geq 2$

En posant $\theta = \text{Arc tg} \frac{t}{\sqrt{n}}$ on a :

n impair :

$$P(|T_n| < t) = \frac{2}{\pi}\left\{\theta + \sin\theta\left[\cos\theta + \frac{2}{3}\cos^3\theta + \cdots + \frac{2.4.....(n-3)}{1.3.....(n-2)}\cos^{n-2}\theta\right]\right\}$$

n pair : $$P(|T_n| < t) = \sin\theta\left\{1 + \frac{1}{2}\cos^2\theta + \cdots + \frac{1.3.5\ldots(n-3)}{2.4.6\ldots(n-2)}\cos^{n-2}\theta\right\}$$

en particulier on en déduit :

$$P(|T_2| < t) = \frac{t}{\sqrt{2 + t^2}} \qquad \text{et} \qquad P(|T_4| < t) = \frac{6t + t^3}{(4 + t^2)^{3/2}}.$$

C.4.2 Formule approchée

Elle se déduit de l'approximation de la loi de Fisher-Snedecor car $T_n^2 = F(1\,;n)$

$$P(|T| > t) \simeq P\left(U > \frac{t^{2/3}\left(1 - \frac{2}{9n}\right) - \frac{7}{9}}{\sqrt{\frac{2}{9} + t^{4/3}\frac{2}{9n}}}\right)$$

D Les fonctions eulériennes Γ et B

D.1 LA FONCTION Γ

Elle est définie pour $x > 0$ par :

$$\boxed{\Gamma(x) = \int_0^{\infty} \exp(-t)t^{x-1}\mathrm{d}t}$$

Relation fondamentale : $\boxed{\Gamma(x+1) = x\Gamma(x)}$

En effet : $$\Gamma(x+1) = \int_0^{\infty} \exp(-t)t^{x}\,\mathrm{d}t = \int_0^{\infty} d(\exp(-t))t^{x}$$

$$= \Big[-\exp(-t)t^{x}\Big]_0^{\infty} + x\int_0^{\infty} \exp(-t)t^{x-1}\,\mathrm{d}t$$

Or $\exp(-t)t^{x}$ vaut 0 si x vaut 0 ou ∞.

On a : $$\Gamma(1) = \int_0^{\infty} \exp(-t)\,\mathrm{d}t = 1$$

d'où : $$\Gamma(n+1) = n\Gamma(n) = n(n-1)\Gamma(n-2) = n!\Gamma(1) = n!$$

$$\boxed{\Gamma(n+1) = n!}$$

La fonction Γ généralise la notion de factorielle aux nombres réels positifs (fig. D.1) :

Lorsque $x \to 0$, $\Gamma(x) \to \infty$.

En effet, supposons $\Gamma(x) \to m$ fini, d'après $\Gamma(x+1) = x\Gamma(x)$ on obtient par continuité si $x \to 0$ $\Gamma(1) = 0$ ce qui est absurde, donc $\Gamma(x) \to \infty$.

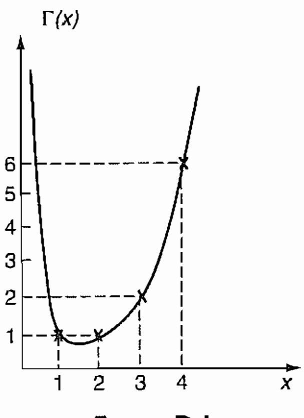


FIGURE D.1

On démontre la formule de Stirling à partir de l'étude de la fonction Γ :

$$n! \simeq n^n \exp(-n)\sqrt{2\pi n}$$

On a aussi la formule : $\Gamma(z)\Gamma(1-z) = \dfrac{\pi}{\sin \pi z}$ et $\Gamma\left(k+\dfrac{1}{2}\right) = \dfrac{1.3.5\ldots(2k-1)}{2^k}\Gamma\left(\dfrac{1}{2}\right)$.

En effet $\Gamma\left(k+\dfrac{1}{2}\right) = \left(k-\dfrac{1}{2}\right)\Gamma\left(k-\dfrac{1}{2}\right) = \dfrac{2k-1}{2}\Gamma\left(k-\dfrac{1}{2}\right)$ d'où le résultat en itérant.

D.2 LA FONCTION *B* DE DEUX VARIABLES

Définition :

$$B(p,q) = \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}$$

en particulier si p et q sont entiers :

$$B(p,q) = \frac{(p-1)!(q-1)!}{(p+q-1)!} = \frac{(p-1)!(q-1)!}{(p+q-1)(p+q-2)!} = \frac{1}{(p+q-1)C_{p+q-2}^{p-1}}$$

Cherchons à exprimer B par une intégrale :

$$\Gamma(p) = \int_0^{\infty} \exp(-t)t^{p-1}\,dt = 2\int_0^{\infty} \exp(-u^2)u^{2p-1}\,du \quad \text{avec} \quad t = u^2$$

donc : $$\Gamma(p)\Gamma(q) = 4\int_0^{\infty}\int_0^{\infty}\exp(-u^2)u^{2p-1}\,\mathrm{d}u\,\exp(-v^2)v^{2q-1}\,\mathrm{d}v$$

$$= 4\int\int\exp(-(u^2+v^2))u^{2p-1}v^{2q-1}\,\mathrm{d}u\,\mathrm{d}v$$

Passons en polaires : $u = \rho\cos\theta \qquad v = \rho\sin\theta$:

$$\Gamma(p)\Gamma(q) = 4\int_{\rho=0}^{\infty}\int_{\theta=0}^{\pi/2}\exp(-\rho^2)\rho^{2p-1+2q-1}(\cos\theta)^{2p-1}(\sin\theta)^{2q-1}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta$$

$$\Gamma(p)\Gamma(q) = 4\int_{\rho=0}^{\infty}\int_{\theta=0}^{\pi/2}\exp(-\rho^2)\rho^{2(p+q)-1}(\cos\theta)^{2p-1}(\sin\theta)^{2q-1}\,\mathrm{d}\rho\,\mathrm{d}\theta$$

$$= 4\int_{\rho=0}^{\infty}\exp(-\rho^2)\rho^{2(p+q)-1}\,\mathrm{d}\rho\int_{\theta=0}^{\pi/2}(\cos\theta)^{2p-1}(\sin\theta)^{2q-1}\,\mathrm{d}\theta$$

$$= 2\Gamma(p+q)\int_{\theta=0}^{\pi/2}(\cos\theta)^{2p-1}(\sin\theta)^{2q-1}\,\mathrm{d}\theta$$

donc : $$\boxed{B(p,q) = 2\int_0^{\pi/2}(\cos\theta)^{2p-1}(\sin\theta)^{2q-1}\mathrm{d}\theta}$$

En particulier : $$B\left(\frac{1}{2},\frac{1}{2}\right) = \frac{\left[\Gamma\left(\frac{1}{2}\right)\right]}{\Gamma(1)} = \left[\Gamma\left(\frac{1}{2}\right)\right]^2 = 2\int_0^{\pi/2}\mathrm{d}\theta = \pi$$

donc : $$\boxed{\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}}$$

En passant en coordonnées cartésiennes, donc en posant $\cos^2\theta = t$ on trouve :

$$\boxed{B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,\mathrm{d}t}$$

E Quelques résultats utiles d'algèbre linéaire

E est un espace vectoriel de dimension finie muni d'une métrique $\mathbf{M}$ (matrice symétrique définie positive). La plupart des propriétés suivantes seront énoncées sans démonstration.

E.1 MATRICES M-SYMÉTRIQUES

Soit $\mathbf{A}$ une matrice carrée n, n. Le produit scalaire dans E étant défini par $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}'\mathbf{M}\mathbf{v}$, l'adjointe $\mathbf{A}^*$ de $\mathbf{A}$ est définie par :

$$\langle \mathbf{A}^*\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{A}\mathbf{v} \rangle \ \forall \mathbf{u}, \mathbf{v}$$

Si $\mathbf{A}^* = \mathbf{A}$ on dit que $\mathbf{A}$ est M-symétrique, ceci entraîne que :

$$\mathbf{u}'\mathbf{M}\mathbf{A}\mathbf{v} = \mathbf{u}'\mathbf{A}'\mathbf{M}\mathbf{v} \ \forall \mathbf{u}, \mathbf{v}$$

donc :

$$\mathbf{M}\mathbf{A} = \mathbf{A}'\mathbf{M}$$

On montre que $\mathbf{A}$ est alors diagonalisable, que ses valeurs propres sont réelles et que ses vecteurs propres sont M-orthogonaux deux à deux, ce qui généralise les propriétés des matrices symétriques.

Si $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ forment une base M-orthonormée de E alors $\sum_{i=1}^{n} \mathbf{u}_i\mathbf{u}_i' = \mathbf{M}^{-1}$.

E.2 PROJECTEURS M-ORTHOGONAUX

Étant donné un sous-espace W de E, $\mathbf{P}$ est la matrice de projection M-orthogonale sur W si $\mathbf{P}\mathbf{y} \in W$ et si $\langle \mathbf{P}\mathbf{y}, \mathbf{y} - \mathbf{P}\mathbf{y} \rangle = 0$ (fig. E.1).

Ce qui revient à écrire que $\mathbf{P}\mathbf{y} \in W \ \forall \mathbf{y}$, que $\mathbf{P}^2 = \mathbf{P}$ et que $\mathbf{P}'\mathbf{M} = \mathbf{M}\mathbf{P}$.

Un projecteur M-orthogonal est une matrice idempotente et M-symétrique.

Les valeurs propres de $\mathbf{P}$ sont alors 1 ou 0 et Trace $\mathbf{P} = \dim W = \text{rang } \mathbf{P}$.

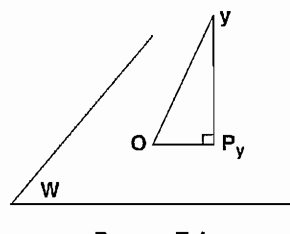


FIGURE E.1

Si $\mathbf{P}_1 + \mathbf{P}_2 + \cdots + \mathbf{P}_k$ sont des projecteurs **M**-orthogonaux alors $\mathbf{P}_1 + \mathbf{P}_2 + \cdots + \mathbf{P}_k$ n'est un projecteur **M**-orthogonal que si, et seulement si, $\mathbf{P}_i\mathbf{P}_j = 0$ pour $i \neq j$, c'est-à-dire si les espaces d'arrivée des $\mathbf{P}_i$ sont **M**-orthogonaux.

Si $W^\perp$ est le supplémentaire **M**-orthogonal de W dans E, alors $\mathbf{I} - \mathbf{P}$ est le projecteur **M**-orthogonal sur $W^\perp$.

Écriture explicite du projecteur P

Supposons W engendré par p vecteurs linéairement indépendants $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ et soit $\mathbf{X}$ la matrice (n, p) ayant les $\mathbf{x}_i$ pour vecteurs-colonnes.

$\mathbf{y} - \mathbf{Py}$ doit être orthogonal à tout vecteur de W ; or, tous les vecteurs de W sont de la forme $\mathbf{Xu}$, en particulier $\mathbf{Py} = \mathbf{Xb}$.

Il faut donc $\langle \mathbf{Xu}_i ; \mathbf{y} - \mathbf{Py} \rangle = 0 \qquad i = 1, 2, \ldots, n \qquad$ où les $\mathbf{u}_i$ forment une base de $\mathbb{R}^p$:

$$\mathbf{u}_i'\mathbf{X}'\mathbf{M}(\mathbf{y} - \mathbf{Py}) = 0 \qquad \forall i$$

donc $\mathbf{X}'\mathbf{My} = \mathbf{X}'\mathbf{MPy}$; comme $\mathbf{Py} = \mathbf{Xb}$ et que $\dim W = p$, $\mathbf{X}'\mathbf{MX}$ est inversible, il vient :

$$\mathbf{X}'\mathbf{My} = \mathbf{X}'\mathbf{MXb} \qquad \text{et} \qquad \mathbf{b} = (\mathbf{X}'\mathbf{MX})^{-1}\mathbf{X}'\mathbf{My}$$
$$\mathbf{Py} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}'\mathbf{MX})^{-1}\mathbf{X}'\mathbf{My}$$

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{MX})^{-1}\mathbf{X}'\mathbf{M}$$

En particulier, le projecteur **M**-orthogonal sur un vecteur $\mathbf{x}$ s'écrit :

$$\mathbf{P} = \mathbf{x}(\mathbf{x}'\mathbf{Mx})^{-1}\mathbf{x}'\mathbf{M} = \frac{\mathbf{xx}'\mathbf{M}}{(\mathbf{x}'\mathbf{Mx})}$$

car $\mathbf{x}'\mathbf{Mx}$ est un scalaire.

E.3 PSEUDO-INVERSES

Soit **A** une matrice rectangle appliquant un espace E dans un espace F. Une matrice $\mathbf{A}^-$ appliquant F dans E telle que :

$$\mathbf{A}^-\mathbf{y} = \mathbf{x} \qquad \text{et} \qquad \mathbf{Ax} = \mathbf{y} \qquad \forall \mathbf{y} \in \mathrm{Im}(\mathbf{A})$$

est appelée pseudo-inverse de **A**.

Il existe toujours au moins un pseudo-inverse qui vérifie la relation caractéristique :

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$$

Il y a en général une infinité de pseudo-inverses, mais il n'existe qu'un seul pseudo-inverse $\mathbf{A}^{+}$, dit de Moore-Penrose, vérifiant en plus :

$$\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$$
$$\mathbf{A}\mathbf{A}^{+} = (\mathbf{A}\mathbf{A}^{+})'$$
$$\mathbf{A}^{+}\mathbf{A} = (\mathbf{A}^{+}\mathbf{A})'$$

Pseudo-inverse d'une matrice diagonale :

$$\mathbf{D} = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

il vient :
$$\mathbf{D}^{+} = \begin{bmatrix} 1/d_1 & & & \\ & 1/d_2 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

Il est alors immédiat de trouver le pseudo-inverse de Moore-Penrose d'une matrice symétrique non régulière en travaillant sur la matrice diagonale de ses valeurs propres.

On en déduit la forme générale du pseudo-inverse de Moore de toute matrice rectangle $\mathbf{A}$:

$$\mathbf{A}^{+} = (\mathbf{A}'\mathbf{A})^{+}\mathbf{A}'$$

où $\mathbf{A}'\mathbf{A}$ est symétrique, en particulier si $\mathbf{A}'\mathbf{A}$ est inversible (le rang de $\mathbf{A}$ est égal au nombre de colonnes de $\mathbf{A}$) $\mathbf{A}^{+} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$.

Si $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ engendrent un espace W de dimension $< p$ le projecteur $\mathbf{M}$-orthogonal sur W est alors :

$$\mathbf{X}(\mathbf{X}'\mathbf{M}\mathbf{X})^{-}\mathbf{X}'\mathbf{M}$$

E.4 FORMULES DE DÉRIVATION VECTORIELLE

Soit g une application de l'espace vectoriel $\mathbb{R}^p$ dans $\mathbb{R} | \mathbf{u} \rightarrow g(\mathbf{u})$.

Par définition on a :

$$\frac{\mathrm{d}g}{\mathrm{d}\mathbf{u}} = \begin{bmatrix} \dfrac{\partial g}{\partial u_1} \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial g}{\partial u_p} \end{bmatrix} \quad \text{où} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_p \end{bmatrix}$$

Propriété 1

$$\frac{d(\mathbf{a}'\mathbf{u})}{d\mathbf{u}} = \mathbf{a} \quad \text{si } \mathbf{a} \text{ est un vecteur constant.}$$

En effet : $$\mathbf{a}'\mathbf{u} = \sum_{i=1}^{p} a_i u_i$$

Donc : $$\frac{\partial(\mathbf{a}'\mathbf{u})}{\partial u_i} = a_i$$

Propriété 2

Soit **A** une matrice carrée de taille p :

$$\frac{d(\mathbf{u}'\mathbf{A}\mathbf{u})}{d\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{A}'\mathbf{u}$$

Soit a_{ij} l'élément courant de **A** ; on a :

$$\mathbf{u}'\mathbf{A}\mathbf{u} = \sum_i \sum_j a_{ij} u_i u_j$$

Cherchons $\frac{\partial(\mathbf{u}'\mathbf{A}\mathbf{u})}{\partial u_1}$: les termes contenant u_1 au premier degré sont de deux sortes : ceux provenant de $u_j = u_i$ et ceux provenant de $u_i = u_1$, c'est-à-dire $\sum_{i \neq 1} a_{i1} u_i u_1$ et $\sum_{j \neq 1} a_{1j} u_1 u_j$ dont les dérivées sont $\sum_{j \neq 1} a_{1j} u_j$ et $\sum_{i \neq 1} a_{i1} u_i$ et il faut ajouter $a_{11} u_1$ à chacun.

On a donc : $$\frac{d(\mathbf{u}'\mathbf{A}\mathbf{u})}{d\mathbf{u}} = \begin{bmatrix} \sum_j a_{1j} u_j \\ \sum_j a_{2j} u_j \\ \vdots \\ \sum_j a_{pj} u_j \end{bmatrix} + \begin{bmatrix} \sum_i a_{i1} u_i \\ \sum_i a_{i2} u_i \\ \vdots \\ \sum_i a_{ip} u_i \end{bmatrix}$$

$$\frac{d(\mathbf{u}'\mathbf{A}\mathbf{u})}{d\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{A}'\mathbf{u}$$

En particulier pour une matrice **H**, symétrique **H** = **H**'.

Donc : $$\frac{d(\mathbf{u}'\mathbf{H}\mathbf{u})}{d\mathbf{u}} = 2\mathbf{H}\mathbf{u}$$

Application à la maximisation du quotient de deux formes quadratiques

Soient **A** et **B** deux matrices symétriques de même taille. **B** sera supposée inversible.

Alors le rapport $\frac{\mathbf{u}'\mathbf{A}\mathbf{u}}{\mathbf{u}'\mathbf{B}\mathbf{u}}$ est maximal pour **u** vecteur propre de $\mathbf{B}^{-1}\mathbf{A}$, associé à sa plus grande valeur propre λ_1, λ_1 étant alors la valeur du maximum.

Démonstration : Un extremum de $\frac{\mathbf{u}'\mathbf{A}\mathbf{u}}{\mathbf{u}'\mathbf{B}\mathbf{u}}$ s'obtient en annulant sa dérivée qui vaut :

$$\frac{(\mathbf{u}'\mathbf{B}\mathbf{u})(2\mathbf{A}\mathbf{u}) - (\mathbf{u}'\mathbf{A}\mathbf{u})(2\mathbf{B}\mathbf{u})}{(\mathbf{u}'\mathbf{B}\mathbf{u})^2}$$

Soit :

$$(\mathbf{u}'\mathbf{B}\mathbf{u})\mathbf{A}\mathbf{u} = (\mathbf{u}'\mathbf{A}\mathbf{u})\mathbf{B}\mathbf{u}$$

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{u} = \left(\frac{\mathbf{u}'\mathbf{A}\mathbf{u}}{\mathbf{u}'\mathbf{B}\mathbf{u}}\right)\mathbf{u}$$

u est donc vecteur propre de $\mathbf{B}^{-1}\mathbf{A}$ associé à la valeur propre $\left(\frac{\mathbf{u}'\mathbf{A}\mathbf{u}}{\mathbf{u}'\mathbf{B}\mathbf{u}}\right)$. Le maximum est donc atteint si cette valeur propre est maximale.

Bibliographie

ALLISON, P.D., *Missing data*, Sage Publications, 2001.

ANDERBERG, M.R., *Cluster analysis for applications*, Academic Press, New York, 1973.

ANDERSON, T.W., *An introduction to multivariate statistical analysis*, Wiley, 3e éd., New York, 2003.

ARDILLY, P., *Les techniques de sondage*, Editions Technip, 2006.

BARDOS, M., *Analyse discriminante*, Dunod, 2001.

BARNETT, V., *Interpreting multivariate data*, Wiley, New York, 1981.

BENJAMINI, Y., HOCHBERG, Y. « Controlling the false discovery rate: a practical and powerful approach to multiple testing ». *Journal of the Royal Statistical Society*, *B*, **57**, 289–300, 1995.

BENOIST, D., TOURBIER, Y., GERMAIN-TOURBIER, S., *Plans d'expériences : construction et analyse*, Tec et Doc Lavoisier, 1994.

BENZÉCRI, J.-P. *et al.*, *L'analyse des données*, tome **I** : *la taxinomie*, tome **II** : *l'analyse des correspondances*, 3e éd., Dunod, Paris, 1979.

BENZÉCRI, J.-P., *Histoire et préhistoire de l'analyse des données*, Dunod, Paris, 1983.

BENZÉCRI, J.-P., *La place de l'a priori*, Encyclopedia Universalis, tome 17, 11–23, Paris.

BERNIER, J., ULMO, J., *Éléments de décision statistique*, PUF, Paris, 1973.

BERTIER, P., BOUROCHE, J.-M., *Analyse des données multidimensionnelles*, PUF, Paris, 1975.

BHATTACHARYYA, G.K., JOHNSON, R.A., *Statistical concepts and methods*, Wiley, New York, 1977.

BIRKES, D., DODGE, Y., *Alternative methods of regression*, Wiley, 1993.

BOUROCHE, J.-M., *Analyse des données en marketing*, Masson, Paris, 1977.

BOUROCHE, J.-M., SAPORTA, G., *L'analyse des données*, Collection Que sais-je, PUF, Paris, 1980.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R.A., STONE, C.J. *Classification and regression trees*, Wadsworth,1984.

BRUYNHOOGHE, M., « Classification ascendante hiérarchique de grands ensembles de données. Un algorithme rapide fondé sur la construction de voisinages réductibles ». *Cahiers de l'Analyse des Données* 3, 1, 1978.

BURGES, C.J.C. « A Tutorial on Support Vector Machines for Pattern Recognition ». *Knowledge Discovery and Data Mining*, **2**(2), 1998.

CAILLIEZ, F. « The analytical solution of the additive constant problem ». *Psychometrika*, **48**, 305–310, 1983.

CAILLIEZ, F., PAGÈS, J.-P., *Introduction à l'analyse des données*, Smash, Paris, 1976.

CAPERAA, P., VAN CUTSEM, B., *Méthodes et modèles en statistique non paramétrique*, Dunod, Paris, 1988.

CARROLL, J.D., « Generalisation of canonical analysis to three or more sets of variables », *Proc. Amer. Psy. Assist.*, 227–228, 1968.

CAZES, P. « Quelques méthodes d'analyse factorielle d'une série de tableaux de données. » *La Revue MODULAD*, **31**, 1–31, 2004.

CHATTERJEE, S., PRICE, B., *Regression analysis by example*, Wiley, New York, 1977.

CIBOIS, P., *L'analyse factorielle*, Collection Que sais-je, PUF, Paris, 1983.

CONOVER, W.J., *Practical nonparametric statistics*, 2e ed., Wiley, New York, 1980.

COOK, R.D., WEISBERG, S., *Residuals and influence in regression*, Chapman and Hall, London, 1982.

DAGNELIE, P., *Analyse statistique à plusieurs variables*, Presses agronomiques de Gembloux, 1975.

DAGNELIE, P., *Théories et méthodes statistiques*, tome I, 1973, tome II. Presses Agronomiques Gembloux, 1975.

DAVISON, A.D., HINKLEY, D.V., *Bootstrap methods and their applications*, Cambridge University Press, 1997.

DE FINETTI, B., *Theory of probability*, 2 tomes, Wiley, New York, 1974.

DE JONG, S. « PLS fits closer than PCR », *Journal of Chemometrics*, **7**, 551–557, 1993.

DEHEUVELS, P., *Probabilité, hasard et certitude*, Collection Que sais-je, PUF, Paris, 1982.

DELECROIX, M., *Histogrammes et estimation de la densité*, Collection Que sais-je, PUF, Paris, 1983.

DEROO, M., DUSSAIX, A.-M., *Pratique et analyse des enquêtes par sondage*, PUF, Paris, 1980.

DEVILLE, J.-C., MALINVAUD, E., « Data analysis in official socio-economic statistics » *JRSS*, série A, **146**, 335–361, 1983.

DEVILLE, J.-C., SAPORTA, G., « Correspondence analysis with an extension towards nominal time series », *Journal of Econometrics*, **22**, 169–189, 1983.

DIDAY, E. *et al.*, *Optimisation en classification automatique*, 2 tomes, Inria, Rocquencourt, 1979.

DIDAY, E., LEMAIRE, J., POUGET, P., TESTU, F., *Éléments d'analyse des données*, Dunod, Paris, 1983.

DRAPER, N.R., SMITH, H., *Applied regression analysis*, Wiley, New York, 1966.

DROESBEKE, J.J., FINE, J., SAPORTA, G. (éditeurs), *Plans d'expériences, applications à l'entreprise*, Editions Technip, 1997.

EFRON, B., *The jackknife, the bootstrap and other resampling plans*, SIAM, New York, 1982.

ESCOFIER, B., PAGÈS, J., *Analyses factorielles simples et multiples*, Dunod, 1988.

ESCOUFIER, Y., « New results and new uses in principal components of instrumental variables », *Proc. 42nd Session Int. Stat. Inst.*, 49–152, 1979.

FELLER, W., *An introduction to probability theory and its applications*, 2 vol., Wiley, New York, 1968 et 1971.

FOURGEAUD, C., FUCHS, A., Statistique, Dunod, 2e éd., Paris, 1972.

FREUND, Y., SCHAPIRE R.E. « A decision-theoretic generalization of on-line learning and an application to boosting. » *Journal of Computer and System Sciences*, **55**, 119–139, 1997.

GENTLE, J. *Random number generation and Monte Carlo methods*, Springer, 2003.

GERI, *Analyse des données évolutives*, Editions Technip, 1996.

GIRI, N., *Multivariate statistical inference*, Academic Press, New York, 1977.

GITTINS, R., *Canonical analysis*, Springer-Verlag, New York, 1985.

GNANADESIKAN, R., *Methods for statistical data analysis of multivariate observations*, Wiley, New York, 1977.

GNEDENKO, B. *et al.*, *Méthodes mathématiques en théorie de la fiabilité*, Mir, Moscou, 1972.

GOODMAN, L., KRUSKAL, W., *Measures of association for cross-classifications*, Springer-Verlag, New York, 1979.

GOURIÉROUX, C., MONFORT, A., Statistique et modéles économetriques, Economica, Paris, 1989.

GOWER, J., HAND, D., *Biplots*, Chapman & Hall, 1996.

GREEN, B., *Analyzing multivariate data*, Holt, Rinehart, Winston, New York, 1978.

GREENACRE, M.J., *Theory and application of correspondence analysis*, Academic Press, New York, 1984.

GUTTMAN, L., « The quantification of a class of attributes. A theory and method of scale construction in the prediction of personal adjustment », 319–348 Soc. Sc. Res. Council, New York, 1941.

HAHN, G.J., MEEKER, W.Q., *Statistical intervals*, Wiley, 1991.

HAND, D.J., « Data mining: statistics and more ? », *The American Statistician*, **52**, 112–118, 1998.

HAND, D.J., *Discrimination and classification*, Wiley, London, 1981.

HARTIGAN, *Clustering algorithms*, Wiley, New York, 1975.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J., *Elements of statistical learning*, Springer, 2001.

HUBER, P., *Robust statistics*, Wiley, New York, 1981.

JACQUARD, A., *Les probabilités*, Collection Que sais-je, PUF, Paris, 1974.

JAMBU, M., LEBEAUX, M.O., *Classification automatique pour l'analyse des données*, tome I : *Méthodes et algorithmes*, tome II : *Logiciels*, Dunod, Paris, 1978.

JAUPI, L., *Contrôle de la qualité*, Dunod, 2002.

JOHNSON, N.L., KOTZ, S., *Distribution in statistics* (4 vol.). Wiley, 1969–1972.

KARLIS, D., SAPORTA, G., SPINAKIS, A., « A Simple Rule for the Selection of Principal Components », *Communications in Statistics – Theory and Applications*, **32**, 3, 643–666, 2003.

KENDALL, M.G., *Rank correlation methods*, Griffin, London, 1962.

KENDALL, M.G., STUART, A., *The advanced theory of statistics* (3 vol.), Griffin, London, 1966.

KRUSKAL, J.B., WISH, M., *Multidimensional scaling*, Sage publications, 1978.

KSHIRSAGAR, A.M., *Multivariate analysis*, Marcel Dekker, New York, 1972.

LECOUTRE, J.-P., TASSI, P., *Statistique non paramétrique et robustesse*, Economica, Paris, 1987.

LEHMANN, E.L., *Non parametrics*, Holden Day, San Francisco, 1975.

LEJEUNE, M., *Statistique, la théorie et ses applications*, Springer, 2005.

MAC QUITTY, L.L., « Similarity analysis by reciprocal pairs of discrete and continuous Data », *Educ. Psych. Meas.*, **26**, 825–831, 1966.

MALINVAUD, E., *Méthodes statistiques de l'économétrie*, Dunod, Paris, 1964.

MARCOTORCHINO, J.F., MICHAUD, P., *Optimisation en analyse ordinale des données*, Masson, Paris, 1979.

MARDIA, K.V., KENT, J.T., BIBBY, J.M., *Multivariate analysis*, Academic Press, London, 1979.

MATALON, B., *Épistémologie des probabilités. In : Logique et connaissance scientifique*, Pléiade, Paris, 526–553, 1967.

MATHERON, G., *Estimer et choisir ; essai sur la pratique des probabilités*, Centre de Morpho. Math., École des Mines de Paris, 1978.

MÉTIVIER, M., *Notions fondamentales de la théorie des probabilités*, Dunod, 2e éd., Paris, 1972.

MOREAU, J., DOUDIN, P.A., CAZES, P., *L'analyse des correspondances et les techniques connexes*. Springer, 2000.

NAKACHE, J.P., CONFAIS, J., *Approche pragmatique de la classification*, Editions Technip, 2005.

NAKACHE, J.P., CONFAIS J., *Statistique explicative appliquée*, Editions Technip, 2003.

NEVEU, J., *Bases mathématiques du calcul des probabilités*, Masson, Paris, 1964.

NISHISATO, S., *Analysis of categorical data : dual scaling and its applications*, University of Toronto Press, 1980.

PAPOULIS, A., *Probability, random variables and stochastic processes*, Mc Graw Hill, New York, 1965.

RAMSAY, J.O., « Monotone regression splines in action », Statistical Science, 3, 425-461, 1988.

RAO, C.R., « The use and interpretation of principal components analysis in applied research », *Sankhya*, A **26**, 329–358, 1964.

RAO, C.R., *Linear statistical inference and its applications*, Wiley, 2e ed., New York, 1973.

RÉNYI, A., *Calcul des probabilités*, Dunod, Paris, 1966.

ROBERT, C., *The bayesian choice*, Springer, 2001.

ROTHSCHILD, J.E, STIGLITZ, M., « Increasing risk : I. a definition ». *J. Econ. Theory*, **2**, 225–243, 1970.

ROUX, M., *Algorithmes de classification*, Masson, Paris, 1986.

SAPORTA, G., « Une méthode et un programme d'analyse discriminante sur variables qualitatives », *Premières Journées Internationales, Analyses des données et informatiques*, INRIA, Rocquencourt, 1977.

SCHÖLKOPFF, B., SMOLA, A., MULLER, K.R. « Nonlinear Component Analysis as a Kernel Eigenvalue Problem », *Neural Computation*, **10**, 1299–1319, 1998.

SCHEFFE, H., *The analysis of variance*, Wiley, New York, 1959.

SCHIFFMAN, S., REYNOLDS, M.L., YOUNG, F.W., *Introduction to multidimensional scaling*, Academic Press, New York, 1981.

SILVERMAN, B.W., *Density estimation for statistics and data analysis*, Chapman and Hall, London, 1986.

SOKAL, R.S., SNEATH, P., *Principles of numerical taxonomy*, Freeman, San Francisco, 1963.

TAKEUCHI, K., YANAI, H., MUKHERJEE, B., *The foundations of multivariate analysis*, Wiley Eastern, New Delhi, 1982.

TASSI, P., *Méthodes statistiques*, Economica, Paris, 1985.

TENENHAUS, M. *La régression PLS*, Edtions Technip, 1998.

THIRIA, S., LECHEVALLIER, Y, GASCUEL, O. (éditeurs), *Statistique et méthodes neuronales*, Dunod, 1997.

TILLÉ, Y., *Théorie des sondages*, Dunod, 2001.

TOMASSONE, R., LESQUOY, E., MILLIEZ, C., *La régression*, Masson, Paris, 1983.

TUFFÉRY, S., *Data Mining et statistique décisionnelle*, Editions Technip, 2005.

TUKEY, J., *Exploratory data analysis*, Addison-Wesley, Reading, 1977.

VAPNIK, V., *Statistical Learning Theory*, Wiley, 1998.

VOLLE, M., *Analyse des données*, Economica, 2e éd., Paris, 1981.

Index des noms

Index

Imprimé en France par EUROPE MEDIA DUPLICATION S.A.S.
53110 Lassay-les-Châteaux
N° 15585 - Dépôt légal : Juin 2006